

An Empirical Likelihood-based Local Estimation^{*}

Zhengyuan Gao[‡]

March 27, 2014

Abstract

This paper proposes a local representation for Empirical Likelihood (EL). EL admits the classical local linear quadratic representation by its likelihood ratio property. A local estimator is derived by using the new representation. Consistency, local asymptotic normality, and asymptotic optimality results hold for the new estimator. In particular, when the regularity conditions do not include any differentiability assumption, these asymptotic results are still valid for the local estimator. Simulations illustrate that the local method improves the inference accuracy of EL.

Key Words: Linear quadratic representation, Infinite divisible family, Local asymptotic normality.

JEL Classification: C40

^{*}The author would like to express his appreciation to Peter Boswijk, Kees Jan van Garderen, Yuichi Kitamura, Richard Smith, Kenneth Judd, Paulo Parente, participants in the seminars at University of Warwick, Toulouse School of Economics, Cowles Foundation, University of Amsterdam and participants in Econometric Society World Congress at Shanghai, Winter North American Econometric Society at Atlanta for helpful comments and discussions. All the remaining errors are mine.

[†]The University of Iowa, John Pappajohn Business Building S362, Iowa City, IA 52242-1994, United States. E-mail: gao-zhengyuan@uiowa.edu

[‡]Southwestern University of Finance and Economics, RIEM Building 217, Chengdu, 61000, P.R.China. E-mail: zgao@swufe.edu.cn

1 Introduction

A family of probability measures $\mathcal{E}_\theta = \{P_\theta; \theta \in \Theta\}$ could represent a class of economic models. For a specific parameter θ in $\Theta \in \mathbb{R}^d$, the probability P_θ measures the performance of the corresponding model. A sequence of papers consider how to attain a suitable P_θ by comparing a specified moment restriction function or moment constraint function

$$\int m(x, \theta) dP_\theta(x) = \mathbb{E}_\theta[m(X, \theta)],$$

to its sample counterpart

$$\int m(x, \theta) dP_n(x) = \frac{1}{n} \sum_{i=1}^n m(X_i, \theta),$$

where P_n is the empirical distribution (empirical measure) and $m(x, \theta)$ is a $k \times 1$ vector with $k \geq d$ for given x and θ .¹

Empirical Likelihood (EL) fills in the gap between Generalized Methods of Moments (GMM) and the classical Maximum Likelihood Estimation (MLE) because it can incorporate the moment constraints into the classical likelihood-based framework. Qin and Lawless (1994), Kitamura and Stutzer (1997), and Smith (1997) have shown that the estimators in both EL and GMM-based estimates share many similar statistical features. As a matter of fact, EL estimation with moment constraints has often been recognized as a moment-based estimation method in econometrics. The particular correspondence between \mathcal{E}_θ and $m(X, \theta)$ by EL is given as follows. For n observations, the moment-based EL is:

$$\max_{\theta, p_1, \dots, p_n} \left\{ \prod_{i=1}^n np_i \left| \sum_{i=1}^n p_i m(X_i, \theta) = 0, p_i \geq 0, \sum_{i=1}^n p_i = 1 \right. \right\}.$$

Function $m(X_i, \theta)$ is of main interest in all moment-based estimation methods.

The connection between the moment-based estimation method and maximization of likelihood ratios comes from dual parameters, that is, the parameters in a dual problem. The dual problem in Kitamura and Stutzer (1997) shows an alternative way of incorporating moment constraints from GMM. The moment constraints no longer appear directly in the

¹Although P_θ is indexed by θ , the true distribution of $m(X, \theta)$ does not depend on θ . The notation P_θ can be interpreted as a pseudo measure of $m(X, \theta)$ and the specification of this measure depends on the value of θ . Later, P_θ is called an implied measure.

objective functions as in GMM or other minimum distance methods. The moment constraints, however, are controlled dually by the Lagrangian multiplier in EL and then appear indirectly in the modified objective functions.² Using the auxiliary dual parameters, Smith (1997) and Newey and Smith (2004) show that a class of estimators including Exponential Tilting, continuous updating GMM and EL, will have better statistical properties than the original GMM whose weighting matrices are not necessarily optimal. However, the minimax type nonlinear optimization induced by the dual parameters makes EL and its related methods less applicable.

The main contributions of the paper are twofold. First, we present a feasible local criterion EL function which resolves the minimax criterion over nonlinear likelihood function. When the likelihood function in the primal problem of EL has nonlinear constraints, the objective function of the dual problem forms a minimax criterion with an infinite dimensional (functional) dual parameter. Without an explicit functional form, the dual parameter cannot be specifically incorporated in a global representation. Furthermore, the dual parameter of EL may have unstable solution(s) that give a thread to estimation and also a thread to computation. Because the dual parameter appears in the criterion function in the primal problem and also appears the Hessian matrix in the optimization algorithm. The localization method will mitigate these threads. The basic idea in this paper is to linearize the nonlinear optimization problem of EL by localizing the likelihood ratio function. Once the nonlinear problem becomes a linearized optimization problem, the minimax problem is reduced to a linear or a quasi-linear programming problem.³

The second contribution is to derive the local estimator, propose its computation method and study its asymptotic properties. The estimator comes from the primal-dual scheme together with Newey-Le Cam's localization. The estimation principle is as follows: approximate the likelihood ratio in the primal problem, obtain a tractable dual representation for the approximating primal problem, update the dual parameter and then return its value to the primal problem. The dual result follows the idea of the Kitamura-Stutzer (Kitamura and Stutzer, 1997) type duality and it assists to adjust the multiplier and the primal likelihood

²Duality theory studies a pair of optimization problems, the initial problem, which refers to the "primal problem", and the dual problem. The aim of dual problem is to obtain more information about the primal problem. For EL and its related methods, the information of constraints and the information of optimal "weights" $\{p_i\}_{i \leq n}$ of these constraints are presented in a single criterion by the duality theory.

³In optimization, when one attempts to solve a nonlinear optimization problem, one should first think about transferring the problem into a linear or quasi-linear environment.

function. Statistical properties of this iterative scheme will depend only on the last iteration of the constructed estimator. This estimator is asymptotically optimal. In addition, the local estimator does not require a differentiable condition of the likelihood function. This result could be important to practitioners. It provides a theoretical ground for the practical use of EL estimator for data with contaminated moment constraints which will be illustrated in Monte Carlo simulations.

In particular, localization representation avoids poor behaviors of likelihood ratios in some corrupted models by contamination. In our consideration, contamination induces non-informative likelihood ratio values for estimation or poorly behaved Hessian matrices for computation. For example, if the likelihood is flat in a neighborhood of some critical points, the Hessian matrix is (near-) singular and the computation may break down at these points. In the implementation, the likelihood of EL includes a vector of implied probabilities $(\tilde{p}(X_1, \theta), \dots, \tilde{p}(X_n, \theta))$ where $\theta \in \Theta \subset \mathbb{R}^d$. Localization considers the probability vector $(\tilde{p}(X_1, \theta^* + \delta_n \tau), \dots, \tilde{p}(X_n, \theta^* + \delta_n \tau))$ on a neighborhood of some θ^* and returns numbers for each τ instead of functions. A well-behaved local representation ensures the existence of the derivative of this representation. By definition, when the derivative exists, small changes will not blow up the approximation of the original likelihood ratio function and this representation is therefore robust to these changes. Thus localization avoids the peculiar points that break down the computational routines.

One could think of this local representation as an alternative criterion function to the likelihood ratio. The following discusses the connection between frequently used criterion functions and the local approximating likelihood ratio criterion in this paper. EL has been embedded into several general criteria, see e.g. Smith (1997), Baggerly (1998), Newey and Smith (2004). The aims of these estimation methods are similar: to optimize a criterion function of θ , such as a likelihood ratio function, subject to some constraint of $m(X, \theta)$. The choice of criterion functions matters for the efficiency and the robustness of an estimator. To balance the tradeoff between these two objectives, Schennach (2007) suggests a two-step inference method by switching the empirical discrepancy between two criterion functions, Kullback-Leibler and likelihood ratio. Although this two-step inferential method works better than either its criterion functions, changing the criterion function in the intermediate stage could distort the supports of likelihood ratio and of Kullback-Leibler functions.⁴ In-

⁴Kullback-Leibler and likelihood ratio use different measures as their dominating measures in the criterion functions. Switching the position of these measures require a mutual contiguity between the empirical

stead of using two-step method, Kitamura et al. (2009) suggest using Hellinger's distance as the criterion. Hellinger's distance has a better topological structure than likelihood ratio and its estimator shares almost the same first order statistical properties with EL. In this paper, our representation of the classical likelihood ratio is a linear-quadratic type approximation. This representation locally obtains some Gaussian properties and therefore maintains a similar topological structure as Hellinger's distance.⁵ The linear-quadratic representation induces the Newton type iteration which is easier for implementations than previous methods since it does not calculate the Hessian based on the second derivative of moment constraints.

The rest of the paper is organized as follows. Section 2 describes EL and gives a version of consistency result without requiring the existence of derivatives. Section 3 presents the local representation of EL. Section 4 gives the local estimator and its asymptotic properties. In Section 5 we describe two Monte Carlo experiments based on linear and nonlinear moment constraints. Finally, conclusions appear in Section 6. Proofs are given in the Appendix.

2 Empirical Likelihood

EL considers a finite dimensional parameter θ and an increasing number of

$$\mathbf{p}(X, \theta) := (p(X_1, \theta), \dots, p(X_n, \theta)).$$

In this paper, the random variable X_i is assumed to be i.i.d.. EL simultaneously finds the optimal θ and the optimal $\mathbf{p}(X, \theta)$ that satisfy the required moment constraints

$$\sum_{i=1}^n m(X_i, \theta) p(X_i, \theta) = 0.$$

Its criterion is:

$$\sup_{p_i, \theta} \left\{ \sum_{i=1}^n \log np_i \mid p_i \geq 0, \quad \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i m_i(\theta) = 0 \right\},$$

measure P_n and implied probability measure \tilde{P}_n . In other words, for every sequence $\{A_n\}_{n \in \mathbb{Z}}$, $P_n(A_n) \rightarrow 0$ implies $\tilde{P}_\theta(A_n) \rightarrow 0$, vice versa. This is a rather strong requirement even for a linear constraint problem.

⁵The covariance function of the approximating log-likelihood ratio process can be attached to an inner product space (pre-Hilbert space) which is close to the L^2 structure considered by the Hellinger distance.

where p_i is a shorthand for $p(X_i, \theta)$ given the value θ . An explicit expression for the optimal p_i 's can be derived using the Lagrangian method and gives the solution:

$$\tilde{p}_i(\theta) := \frac{1}{n} \frac{1}{1 + \lambda_n^T m_i(\theta)},$$

where $\tilde{p}_i(\theta)$ is called the *implied probability*. The candidate solutions belong to the family

$$\mathcal{E}_\theta := \{\tilde{P}_\theta : \theta \in \Theta, \int m(X, \theta) d\tilde{P}_\theta = 0, \tilde{P}_\theta \ll P_0, \tilde{P}_\theta \ll P_n\},$$

where $d\tilde{P}_\theta(x_i) = \tilde{p}_i(\theta) d\mu$ for a counting measure μ .⁶ $\{\tilde{P}_\theta \ll P_0, \tilde{P}_\theta \ll P_n\}$ means that \tilde{P}_θ is contiguous with respect to both P_0 and P_n . For every sequence $\{A_n\}_{n \in \mathbb{Z}}$, $P_n(A_n) \rightarrow 0$ implies that $\tilde{P}_\theta(A_n) \rightarrow 0$ and meanwhile $P_0(A_n) \rightarrow 0$ implies that $\tilde{P}_\theta(A_n) \rightarrow 0$. The λ_n is the solution of:

$$\frac{1}{n} \sum_{i=1}^n \left[\frac{m_i(\theta)}{1 + \lambda_n^T m_i(\theta)} \right] = 0. \quad (2.1)$$

Let the average log-likelihood ratio of the implied probability between any two parameter values θ_1 and θ_2 be:

$$\Lambda_n(\theta_1, \theta_2) := \frac{1}{n} \sum_{i=1}^n \log \left[\frac{\tilde{p}_i(\theta_1)}{\tilde{p}_i(\theta_2)} \right]$$

and define the average log-likelihood ratio of the implied probability given θ and counting numbers $1/n$ as

$$\Lambda_n(\theta) := \frac{1}{n} \sum_{i=1}^n \log n \tilde{p}_i(\theta).$$

The constraint $0 \leq \tilde{p}_i \leq 1$ requires that the inequality $1 + \lambda_n^T m_i(\theta) \geq 1/n$ always holds. The population $\lambda(\theta) := \lim_{n \rightarrow \infty} \lambda_n$ must lie in a convex and closed set $\Gamma_\theta = \lim_{n \rightarrow \infty} \cup_{i=1}^n \Gamma_{\theta,i}$. For fixed n , the set $\Gamma_{\theta,n}$ is defined as a collection of subsets of

$$\{\lambda_n : 1 + \lambda_n^T m(X_i, \theta) \geq 1/n, i = 1, \dots, n, \theta \in \Theta\}.$$

In the rest of this section, we derive an other consistency result for EL estimation. Our intention is to obtain the consistency result without assuming the differentiability of the moment restriction $m(X, \theta)$. The differentiability is often assumed because it is a natural

⁶The family \mathcal{E}_θ obtains both continuous measures and discrete measures. The definition will become clear once we introduce the infinite divisibility concept.

way to derive an expansion of the objective function at the true parameter. This expansion will link the asymptotic behaviors of $T_n - \theta_0$ with those of the sample averages of $\partial m(X, \theta)/\partial \theta$ and hence it is useful for proving both strongly and weakly convergences. But as a trade-off, one needs to impose additional identification conditions and limit distribution conditions for $n^{-1} \sum_i^n \partial m(X_i, \theta)/\partial \theta$ and $n^{-\frac{1}{2}} \sum_i^n \partial m(X_i, \theta)/\partial \theta$ respectively. Because our representation will not rely on such an expansion, we weaken the conditions for consistency.

There are many existing results of EL's consistency. Kitamura et al. (2004) relax the assumptions in Qin and Lawless (1994) and Kitamura and Stutzer (1997) and obtain consistency of the estimator based on Wald's approach (Wald, 1949). Newey and Smith (2004) assume the differentiability of Lagrangian multiplier rather than that of $m(X, \theta)$. However, due to the non-analytical form of $\lambda(\theta)$, this assumption is quite strong. Schennach (2007) gives another consistency proof for a non-differentiable objective function and avoids applications of a Taylor expansion. The differentiability of the moment restriction, however, is still assumed there in order to obtain a valid approximation for the Lagrangian $\lambda(\theta)$. The conditions in the following Theorem 1 are similar to the standard M -estimator conditions in Huber (1981), thus the differentiability assumption is not required. In order to ensure the EL estimator consistent for this case, we need to give a result of EL consistency under weaker conditions. Here are the conditions:

Condition 1. (i) $M(\theta) := \mathbb{E}[m(X, \theta)]$ exists for all $\theta \in \Theta$ and has a unique zero at $\theta = \theta_0$.

(ii) θ_0 is a well-separated point in $M(\theta)$ such that

$$\inf_{\theta: d(\theta, \theta_0) \geq \epsilon} |M(\theta)| > |M(\theta_0)| = 0,$$

where ϵ is an arbitrary value larger than zero and $d(\cdot, \cdot)$ is any distance function on $\Theta \times \Theta$.

(iii) $m(X, \theta)$ is continuous in θ ,

$$\lim_{\theta' \rightarrow \theta} \|m(X, \theta) - m(X, \theta')\| = 0.$$

(iv) Let ∞ be the one-point compactification of Θ , then there exists a continuous function $b(\theta)$ bounded away from zero, such that (1) $\sup_{\theta \in \Theta} m(X, \theta)/b(\theta)$ is integrable, (2) $\liminf_{\theta \rightarrow \infty} \|M(\theta)\|/b(\theta)$ is larger than 1, and (3) $\limsup_{\theta \rightarrow \infty} \|m(X, \theta) - M(\theta)\|/b(\theta) < 1$.

(v) $\sum_{i=1}^n [m(x_i, \theta_0)m(x_i, \theta_0)^T]/n$ is full rank for all $n \geq 1$.

Condition 1 (i) ensures the model is identified for a small neighborhood of θ_0 . (ii) is a local separability condition. (iii) is used to obtain the continuity of the Lagrangian multiplier. (iv) is an envelope assumption; it is used to obtain some dominated convergence results. The one-point (Alexandroff) compactification allows us to let θ approach any boundary place of Θ , even if Θ is not compact and may extend indefinitely. The usual proof of EL consistency (Qin and Lawless, 1994) requires the existence of the continuous derivative of $m(X, \theta)$ and that the derivative is of full rank. Condition 1 is less restrictive because it allows for irregular cases where the usual “delta method” does not work, e.g. when $m(X, \theta)$ is non-differentiable. Condition 1 (i)-(iv) are the standard M-estimator conditions in Huber (1981) and are very weak in the context of parametric models.

Theorem 1. *If Condition 1 holds, then every sequence T_n satisfying*

$$T_n := \arg \sup_{\theta \in \Theta} \sum_{i=1}^n \log n \tilde{p}_i(\theta) = \arg \sup_{\theta \in \Theta} n \Lambda_n(\theta)$$

will converge to θ_0 almost surely.

Note that this theorem does not require any differentiation condition. However, the differentiability is implicitly obtained in the later section. In fact, the “local” concept is the analog of “differential”. If one fixes a particular θ_0 in Θ and investigates what happens to the likelihood ratio function with parameter sequences of the form $\theta = \theta_0 + \delta_n \tau$, with $\delta_n \rightarrow 0$ as n goes to infinity, then δ_n yields a sort of differentiation rate just as the differentiation rate in basic calculus, and then the whole localization problem can be analyzed as a kind of differentiability analysis for the likelihood ratio function. The term τ is called local parameter since it is an index for local features. This technique often appears in the evaluation of local power of test statistics and statistical experiments, see van der Vaart (1998) and Le Cam and Yang (2000).

3 Gaussian Properties and Localization of EL

A non-closed form dual parameter λ_n induces a non-closed form probability vector $\tilde{\mathbf{p}}(\theta)$. General techniques such as empirical processes of studying irregular behavior of the functions are also not directly applicable because the functional form of λ_n has no closed-form representation, since it is the solution of Equation (2.1) that depends on the sample size and

parameter values. In this section, we propose alternative conditions and specifications of EL to standardize the problem.

3.1 Approximation for an Infinitely Divisible Family

Instead of studying the implied probability vectors $\tilde{\mathbf{p}}(\theta)$, we consider a family of probability measures

$$\mathcal{E}_\theta := \{\tilde{P}_\theta : \theta \in \Theta, \int m(X, \theta) d\tilde{P}_\theta = 0\},$$

where the discrete vector $\tilde{\mathbf{p}}(\theta)$ satisfies

$$\sum_i^n m(X_i, \theta) \tilde{p}_i(\theta) = 0.$$

If a random variable ξ , for every natural number n , can be represented as the sum

$$\xi = \xi_{1,n} + \xi_{2,n} + \cdots + \xi_{n,n}$$

of n i.i.d random variables $\xi_{1,n}, \dots, \xi_{n,n}$, then ξ is called *infinitely divisible* (Gnedenko and Kolmogorov, 1968, p. 78). A probability distribution is said to be infinitely divisible if and only if it can be represented as the distribution of the sum of an arbitrary number of i.i.d random variables. A family of such distributions is often referred to as an *infinitely divisible family*. In our case, for arbitrary sample size n and fixed θ , the log-likelihood ratio process is

$$\Lambda((X_1, \dots, X_n), \theta) = \log n\tilde{p}(X_1, \theta) + \cdots + \log n\tilde{p}(X_n, \theta).$$

Every additional term $\log n\tilde{p}(X_i, \theta)$ is an identical distributed increment of this log-likelihood ratio process. One crucial deficiency of the above argument for EL is that $\tilde{p}(X_i, \theta)$ are not independent for all i s. Because λ_n appears in $\tilde{p}(X_i, \theta)$ for $i = 1, \dots, n$. But since the dependence is introduced by λ_n only and λ_n appears as the same form for all $\tilde{p}(X_i, \theta)$, once the value of λ_n is conditioning, the rest part of $\tilde{p}(X_i, \theta)$ will be independent with $\tilde{p}(X_j, \theta)$ for any $i \neq j$.

For a sufficient large n and a fixed θ , λ_n in $\log n\tilde{p}(X_1, \theta)$, is a stochastic element.⁷ In this case, one can think that the integral of the log-likelihood ratio process, $\sum_i \log n\tilde{p}(X_i, \theta)$,

⁷In a localization approach, when θ is given, λ_n will converge to a normal random variable with mean zero, see e.g. Theorem 1 in Qin and Lawless (1994).

represents an infinite divisible process ξ in n additive terms $\xi_{1,n} + \xi_{2,n} + \dots + \xi_{n,n}$.⁸ Thus \mathcal{E}_θ does not merely include the family of distributions that satisfy the constraint $\int m(x, \theta) d\tilde{P}_\theta(x)$, it also requires the sample average of the log-likelihood ratio process of \tilde{P}_θ to be infinitely divisible. It seems that EL inherits the moment constraint from moment-based methods and inherits the infinitely divisibility from likelihood ratio based methods.

An infinitely divisible family \mathcal{E} admits a representation $\mathcal{E} = \mathcal{E}_1 \times \dots \times \mathcal{E}_n = \otimes_{i=1, \dots, n} \mathcal{E}_i$ based on n copies of the so called divisor \mathcal{E}_i , where n could be arbitrarily large and \times denotes the direct product. The family \mathcal{E} is called *divisible* with divisor \mathcal{E}_i . There are several well known infinitely divisible families, e.g. Poisson and Gaussian families.

It has been proved by Gnedenko and Kolmogorov (1968, Theorem 17.5) that any infinitely divisible family can be approximated by a finite number of Poisson type measures. This result basically means that the infinitely divisible family constructed by $\{\log n\tilde{p}(X, \theta)\}$ can be approximated by a finite number of Poisson measures.⁹ Poisson family relates to the Gaussian family via the Hellinger's affinity. We will use this property to deduce a representation of the likelihood ratio process.

Theorem 2. *If \tilde{P}_θ is infinitely divisible then when $n \rightarrow \infty$, the log-likelihood $\log d\tilde{P}_{\theta+\delta_n\tau_n}/d\tilde{P}_\theta$ can be approximated by a linear quadratic expression such that the difference*

$$\sum_{i=1}^n \log \frac{d\tilde{P}_{\theta+\delta_n\tau_n}}{d\tilde{P}_\theta}(x_i) - \left[\tau_n^T S_{\theta,n} - \frac{1}{2} \tau_n^T K_{\theta,n} \tau_n \right] \quad (3.1)$$

tends to zero in probability for any bounded sequence $\{\tau_n\}$ with a random vector $S_{\theta,n}$ and a deterministic matrix $K_{\theta,n}$.

The infinite divisible feature gives us a useful representation for the likelihood ratio process, a linear quadratic expression with a local parameter τ_n . This representation is similar as the linerization method based on Taylor's expansion, however it does not require the differentiability of the implied probability. With this expression, we can construct our estimator without bothering with non-linear optimization, since the parameter in (3.1) is re-parametrized by τ_n which appears linearly and quadratically in the equation. Furthermore, neither the computational algorithm nor the weakly convergent statistics involve any

⁸More details about such a construction are discussed in Le Cam and Yang (2000, Chapter 5), although in most cases, they use $\log(1 + (p_\theta/p_\vartheta)^{-1/2} - 1)$ instead of $\log(p_\theta/p_\vartheta)$ directly.

⁹We give a short description about Poissonization in the appendix. Infinite divisible family holds for arbitrary number of n , so the approximation in principle should be valid for the finite many n .

differentiation requirements.

Remark 1. In the proof, we will show a relation for univariate Gaussian families. For any pair of Gaussian measures G_θ and G_ϑ , there will be a linear-quadratic expression to relate them. Therefore, the integral of $(dG_\theta/dG_\vartheta)^{1/2}$ w.r.t. G_ϑ will have a linear quadratic representation. Then we show that if \tilde{P}_θ is infinitely divisible, $(d\tilde{P}_\theta/d\tilde{P}_\vartheta)^{1/2}$ will be approximately equal to $(dG_\theta/dG_\vartheta)^{1/2}$, so $(d\tilde{P}_\theta/d\tilde{P}_\vartheta)^{1/2}$ will also have a linear quadratic representation.

Remark 2. The linear-quadratic approximations to the log-likelihood ratios can possibly be used with other minimum contrast estimators, but such constructions only lead to asymptotically sufficient estimates, in the sense of Le Cam, when the contrast function mimics the properties of log-likelihood function, at least locally.

Remark 3. From a computational aspect, when confronted with the nonlinear optimization, the Hessian matrix of the problem in some cases is difficult to evaluate especially in regions that are either extremely flat or very erratic. It is then computationally more efficient to consider the local optimization and avoid a singular or non-invertible Hessian matrix rather than calculate the global second order derivative of the objective function.

Remark 4. Theorem 2 shows that with a proper choice of δ_n , the log-likelihood ratio can be approximated by a linear-quadratic representation. One of the main focus of this representation is the quadratic term. For a pair of Gaussian measures (G_θ, G_ϑ) with dominating measure μ we will have

$$\begin{aligned} & \int \left(\frac{dG_\theta}{dG_\vartheta} \right)^{\frac{1}{2}} dG_\vartheta = \int dG_\theta^{\frac{1}{2}} dG_\vartheta^{\frac{1}{2}} d\mu \\ &= \mathbb{E} \exp \left\{ \sum_{i=\theta, \vartheta} \frac{1}{2} \left[L(i) + \mathbb{E} \log \left(\frac{dG_i}{d\mu} \right) \right] \right\} \\ &= \left[\exp -\frac{1}{4} (K(\theta, \theta) + K(\vartheta, \vartheta)) \right] \cdot \mathbb{E} \exp \left(\sum_{i=\theta, \vartheta} \frac{1}{2} L(i) \right) \end{aligned} \quad (3.2)$$

$$= \exp \left\{ \frac{1}{4} [2K(\theta, \vartheta) - K(\theta, \theta) - K(\vartheta, \vartheta)] \right\}, \quad (3.3)$$

where $L(i) := \{\log(dG_i/d\mu) - \mathbb{E} \log(dG_i/d\mu)\}$ for $i = \theta, \vartheta$. The derivation of (3.3) is given in the Appendix. The property of $L(i)$ includes that it is Gaussian with expectation $\mathbb{E}L(i) = 0$ and covariance kernel $K(\theta, \vartheta) = \mathbb{E}L(\theta)L(\vartheta)$ and we have $\mathbb{E}L(i)^2 = K(i, i)$. Let

$$q(\theta, \vartheta) = -8 \log \int dG_\theta^{\frac{1}{2}} dG_\vartheta^{\frac{1}{2}} d\mu.$$

Since the quadratic term is deterministic in the neighborhood of θ_0 , we can use interpolation to find $K(\cdot, \cdot)$. With an arbitrary mid-point u , three-point interpolation gives us:

$$K(\theta, \vartheta) = - (q(\theta, \vartheta) - q(\theta, u) - q(u, \vartheta)).$$

For small $|\theta - \vartheta|$, to speed up the computation, one could use an approximated value $\Lambda_n(\theta, \vartheta)$ instead of $q(\theta, \vartheta)$.¹⁰

3.2 Comparison with Other Conditions

The standard EL ratio can be put into the form of the linear quadratic representation in (3.1) but this requires some additional assumptions, e.g. differentiability of $m(X, \theta)$. The following proposition establishes this relation.

Proposition 1. *Suppose that in addition to Condition 1, the following holds*

(i) *the model is just-identified, $\partial m(X, \theta)/\partial \theta < \infty$ for any X , the rank of $\mathbb{E}[\partial m(X, \theta)/\partial \theta]|_{\theta_0}$ equals $\dim(\theta)$,*

(ii) *$\frac{1}{n} \sum_{i=1}^n [m_i(\theta) m_i(\theta)^T]$ and $\frac{1}{n} \sum_{i=1}^n [\lambda_n^T m_i(\theta)]^2$ are both finite for any positive n , even as $n \rightarrow \infty$,*

then the log-likelihood ratio between \tilde{p}_{θ_0} and $\tilde{p}_{\theta_0 + \delta_n \tau}$ can be approximated by:

$$2 \sum_{i=1}^n \log \frac{\tilde{p}_{\theta_0 + \delta_n \tau_n}(x_i)}{\tilde{p}_{\theta_0}} = \delta_n \tau_n^T A_1 + \frac{1}{2} \delta_n^2 \tau_n^T A_2 \tau_n^T + o_p(1) \quad (3.4)$$

where

$$A_1 = \mathbb{E} \frac{\partial m(X, \theta_0)^T}{\partial \theta} (\mathbb{E} m(X, \theta_0) m(X, \theta_0)^T)^{-1} \sum_{i=1}^n m_i(\theta_0),$$

$$A_2 = \mathbb{E} \frac{\partial m(X, \theta_0)^T}{\partial \theta} (\mathbb{E} m(X, \theta_0) m(X, \theta_0)^T)^{-1} \mathbb{E} \frac{\partial m(X, \theta_0)}{\partial \theta^T}.$$

The expansion (3.4) is obtained simply by Taylor expansion and the result therefore does not apply to the nonstandard problem where the differentiability of $m(X, \theta)$ is questionable. However, the result is intuitive as it mimics the standard Local Asymptotic Normal (LAN)

¹⁰The concern is that the square root density computing may induce rounding error. In fact $\frac{1}{2} \log \int (dG_\theta/dG_\vartheta)^{1/2} dG_\vartheta$ approximately equals to $\frac{1}{2} \sum_i \log(dG_\theta/dG_\vartheta)(x_i)$ when x_i is generated by G_ϑ .

property for parametric models, see e.g. van der Vaart (1998, pp 104). The relation between (3.4) and (3.1) is also quite clear: the first term is τ_n times a random vector, and the second term is its variance.

Remark 5. With the additional normality assumption on the average of $m_i(\theta_0)$ and assuming $\delta_n = n^{-1/2}$ we will of course have:

$$\begin{aligned} & \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbb{E} \frac{\partial m(X, \theta_0)^T}{\partial \theta} (\mathbb{E} m(X, \theta_0) m(X, \theta_0)^T)^{-1} m_i(\theta_0) \\ & \rightsquigarrow \mathcal{N} \left(0, \mathbb{E} \frac{\partial m(X, \theta_0)^T}{\partial \theta} (\mathbb{E} m(X, \theta_0) m(X, \theta_0)^T)^{-1} \mathbb{E} \frac{\partial m(X, \theta_0)}{\partial \theta^T} \right). \end{aligned}$$

Asymptotic normality of the EL estimator is established by equation (3.4) with additional conditions on the continuity or the boundedness of second derivative of the moment restriction functions, e.g. Qin and Lawless (1994), Newey and Smith (2004) or Kitamura et al. (2004).

Remark 6. An alternative way of deducing this asymptotic normality is via Differentiability in Quadratic Mean (DQM). This entails the existence of a vector of measurable functions $S_{\theta_0, n}$ such that

$$\int \left[\tilde{p}_{\theta_0 + \delta_n \tau}^{1/2} - \tilde{p}_{\theta_0}^{1/2} - \frac{1}{2} \delta_n \tau^T S_{\theta_0, n} \tilde{p}_{\theta_0}^{1/2} \right]^2 d\mu = o(\|\delta_n\|^2), \quad (3.5)$$

where $\delta_n \rightarrow 0$. Note that the relation between the derivatives of the square root density and the score function (when it exists) is:

$$2 \frac{1}{\sqrt{\tilde{p}_\theta}} \frac{\partial}{\partial \theta} \sqrt{\tilde{p}_\theta} = \frac{\partial}{\partial \theta} \log \tilde{p}_\theta.$$

If along a path, the square root of the implied probability $\theta \mapsto \sqrt{\tilde{p}_\theta}$ is differentiable, then DQM basically means that a expansion of the square root of \tilde{p}_θ is valid and the remainder term is negligible in $L^2(\mu)$ norm. The term $S_{\theta, n}$ can be considered as the score function of the implied probability \tilde{p}_θ at θ_0 . DQM implies that the condition does not require the point-wise definition of the derivative of $m(\theta, X)$ therefore it is less restrictive.

Suppose the implied probability includes the term $m(\theta, X)$ which is not always differentiable. Then it deserves more efforts to relax the restrictive condition on differentiability. In fact, Theorem 2 implies that the log-likelihood ratio belongs to the LAN family. The result is already good enough for constructing an efficient (or asymptotic sufficient) estimator.

The expression in (3.1) is much weaker than the regular conditions and DQM. It only states that log-likelihood ratios of implied probabilities can be approximated by a linear-quadratic expression.

4 Local Estimation

By the result (3.1) in Theorem 2, we can study the behavior of a pair $(\tilde{P}_{\theta+\delta_n\tau_n}, \tilde{P}_\theta)$ by looking at the log-likelihood ratio process $\Lambda_n(\theta + \delta_n\tau_n, \theta)(X)$ with index τ_n . The log-likelihood ratio process admits linear quadratic approximations as $n \rightarrow \infty$, with the term $\tau_n S_n$ linear in τ_n and the term $\tau_n^T K_n \tau_n$ quadratic in τ_n . The numerical values of the approximation depend on the concentrated point θ and its local neighborhoods. With these ideas in mind, we will show the following steps of constructing a local type estimator. The explanation of each step is given after the definition.

Definition. Given Condition 1, we define the following Le Cam type local EL estimator in 5 steps:

Step 1. Find an auxiliary estimate θ_n^* using a δ_n -consistent estimator and restricted such that it lies in Θ_n (a δ_n -sparse discretization of Θ).

Step 2. Construct a matrix K_n with $K_{n,i,j} = u_i^T K_n u_j$, $i, j = 1, 2, \dots, d$, given by

$$K_{n,i,j} = - \{ \Lambda_n[\theta_n^* + \delta_n(u_i + u_j), \theta_n^*] \\ - \Lambda_n[\theta_n^* + \delta_n u_i, \theta_n^*] - \Lambda_n[\theta_n^* + \delta_n u_j, \theta_n^*] \}$$

and $\{u_1, \dots, u_d\}$ is a set of directional vectors in \mathbb{R}^d . u_i is a step-size in selected in advance.

Step 3. Construct the linear term:

$$u_j^T S_n = \Lambda_n[\theta_n^* + \delta_n u_j, \theta_n^*] + \frac{1}{2} K_{n,j,j}.$$

Since all the right hand side values are known, S_n can be computed and is a proper statistic.

Step 4. Construct the adjusted estimator:

$$T_n = \theta_n^* + \delta_n K_n^{-1} S_n.$$

Step 1 The δ_n -sparse (discretization of the) parameter space in Step 1 is suggested by Le Cam (see Le Cam and Yang (2000, p 125)). It requires a sequence of subsets $\Theta_n \subset \Theta$ satisfying the following conditions (i) that for any $\theta \in \Theta$ and any constant $b \in \mathbb{R}^+$, the ball $B(\theta, b\delta_n)$ contains a finite number of elements of Θ_n , independent of n , and (ii) that there exist a $c \in \mathbb{R}^+$ such that any $\theta \in \Theta$ is within a distance $c\delta_n$ of a point of Θ_n . If we think of Θ_n as nodes of a grid with a mesh that gets finer as n increases, then (i) says that the grid does not get too fine too fast and (ii) says that the mesh refines fast enough to have nodes close to any point in the original space Θ . In other words, asymptotically θ_n^* should be close enough to θ_0 . Another interpretation of δ_n -sparsity is from a Bayesian perspective. That is for arbitrary priors, the corresponding posteriors essentially concentrate on the small vicinities shrinking at the rate δ_n .

Step 2 As in the Remark 4, the covariance matrix in Step 2 is an analog to the covariance kernel in Gaussian processes. For a stationary Gaussian process, the covariance kernel is smooth and differentiable in quadratic mean, the covariance kernel can be written as

$$\begin{aligned} & \text{Cov} \left(\frac{1}{\delta_n} (G_{\theta+u\delta_n} - G_\theta), \frac{1}{\delta_n} (G_{\vartheta+u\delta_n} - G_\vartheta) \right) \\ &= \frac{1}{\delta_n^2} (2C(\theta - \vartheta) - C(\theta - \vartheta + u\delta_n) - C(\theta - \vartheta - u\delta_n)) \\ &\rightarrow - \left. \frac{\partial^2 C(h)}{\partial h^2} \right|_{h=\theta-\vartheta}, \end{aligned}$$

where $C(\theta, \vartheta) := \text{Cov}\{G_\theta, G_\vartheta\}$. Since K_n is an analog to the covariance kernel, the construction of K_n is nothing else but a finite difference of $\Lambda_n(\cdot, \cdot)$ which is analogous to the second derivative of the covariance kernel.

Step 3 and 4 With a control term K_n which is asymptotically determined, all the randomness of the log-likelihood ratio is contained in the first term, S_n . Step 3 is to extract the randomness from $\Lambda_n(\cdot, \cdot)$ and construct the linear term. Step 4 is to construct the estimator. To verify these two steps, we need to ensure that the covariance kernel in (3.1) is invertible.

Proposition 2. *The matrices $K_{\theta,n}$ in (3.1) are almost surely positive definite. Any cluster point K_θ of $K_{\theta,n}$ in $P_{\theta,n}$ -law is invertible.*

If $K_n - K_{\theta,n}$ converges to zero, then K_n is also invertible. This result will be given in the

following Theorem 3. If K_n is positive definite, by substituting $S_n = K_n \delta_n^{-1} (T_n - \theta_n^*)$ into the linear quadratic expression:

$$\begin{aligned} \tau_n^T K_n \delta_n^{-1} (T_n - \theta_n^*) - \frac{1}{2} \tau_n^T K_n \tau_n = & -\frac{1}{2} \delta_n^{-2} [T_n - (\theta_n^* + \delta_n \tau_n)]^T K_n \times \\ & [T_n - (\theta_n^* + \delta_n \tau_n)] + \frac{1}{2} \delta_n^{-2} [T_n - \theta_n^*]^T K_n [T_n - \theta_n^*], \end{aligned}$$

we have a quadratic expression of T_n and $(\theta_n^* + \delta_n \tau_n)$. The maximal value of this approximating representation of the log-likelihood ratio is achieved when $\theta_n^* + \delta_n \tau_n = T_n$. In other words, $\delta_n^{-1} (T_n - \theta_n^*)$ is the estimator for the local parameter τ_n .

Remark 7. The construction was originally proposed by Le Cam (1974). He supposed that there is a special interest in the likelihood function at particular points where Taylor's expansion fails, e.g. for the Laplace distribution. The advantage of the construction is that the quadratic term does not depend very much on the particular auxiliary estimation method that is used to obtain the value of θ_n^* and the construction is only determined in a local neighborhood of the particular point.

Remark 8. One may be concerned with the δ_n -consistency requirement for the auxiliary estimator. For a simple i.i.d. case, the δ_n is set to $n^{-1/2}$, the requirement is the same as asking for an \sqrt{n} -consistent auxiliary estimator. Any \sqrt{n} -consistent estimator should be, in principle, good enough from the estimation perspective, because the auxiliary estimator θ_n^* is at least in a neighborhood of θ_0 . However, in practice, it may be hard to find a well behaved moment restriction function around θ_0 . The use of local EL estimator is to overcome the problem and improve the auxiliary estimator. We suppose that θ_n^* is located within a range $n^{-1/2}$ of the true value, then a local method would give a refinement. When consistency and asymptotic normality are treated separately, one could take good care of consistency first and then use localization method to improve the final result or one could take care of the concentration of distribution first and then correct the bias by localization.

Theorem 3. *Given Condition 1, T_n , S_n and K_n have following properties:*

- (i) $K_n^{-1} S_n - K_{\theta,n}^{-1} S_{\theta,n}$ and $K_n - K_{\theta,n}$ converge to zero in $\tilde{P}_{\theta,n}$ -law where $(K_{\theta,n}, S_{\theta,n})$ is in (3.1).
- (ii) $\delta_n^{-1} (T_n - \theta)$ is bounded in $\tilde{P}_{\theta,n}$ -law.
- (iii) if Equation (3.5) holds and the moment restrictions are just-identifying, the sequence

of models $\{\tilde{P}_{\theta,n} : \theta \in \Theta\}$ is LAN and

$$\delta_n^{-1}(T_n - \theta_0) \rightsquigarrow \mathcal{N}(0, \Omega)$$

where $\Omega = \mathbb{E} \frac{\partial m(x, \theta_0)}{\partial \theta}^T (\mathbb{E} m(x, \theta_0) m(x, \theta_0)^T)^{-1} \mathbb{E} \frac{\partial m(x, \theta_0)}{\partial \theta}$.

The LAN theory is useful in showing that many statistical models can be approximated by Gaussian models. In the parametric likelihood framework, when the original model P_θ is smooth in the parameters, i.e. DQM, the local parameter $\tau_n = \delta_n^{-1}(\theta_0 - \theta_n^*)$ can be used to construct a log likelihood ratio based on $P_{\theta_0 + \tau_n \delta_n}$ that is asymptotically $\mathcal{N}(\tau_n, I_{\theta_0}^{-1})$. Here we use LAN in a moment based setting without further parametric assumptions. Once LAN is established, asymptotic optimality of estimators and of tests can be expressed in terms of LAN properties.

Remark 9. Some other articles also utilize local information based on an EL framework. Donald et al. (2003) propose resampling data from a local EL estimated distribution. Kitamura et al. (2004) consider another localized EL based on conditional moment restrictions and use them to re-construct a smooth global profile likelihood function. Smith (2005) extends moment smoothing to GEL. These methods construct smooth objective functions, implicitly or explicitly. Our solution is to discretize the parameter space and then construct local log-likelihood ratios as local objective functions. Thus our localization is viewing a different aspect of the problem.

Theorem 3 gives an asymptotic result on the weak convergence of the estimator. In the theorem, the limit distribution is based on a kind of Cramér-Rao type lower bound and is essentially a point-wise result. In order to obtain a result in a neighborhood rather than at a single point, we will now state and prove a minimax type theorem on the risk of any estimator.

Before giving the theorem, we need to introduce a technical concept of δ_n -regularity. This concept expresses the desirable requirement that a small change in the parameter should not change the distribution of estimator too much. For the estimator sequence T_n , if the difference between the distributions of $\delta_n^{-1}(T_n - \theta_0 - \delta_n \tau)$ and $\delta_n^{-1}(T_n - \theta_0)$ tends to zero under $P_{\theta_0 + \delta_n \tau, n}$ -law and $P_{\theta_0, n}$ -law respectively, then T_n is called δ_n -regular at the point θ_0 .

Theorem 4. *Given Condition 1 and letting W be a non-negative bowl shaped loss function,*

if T_n is δ_n -regular on all Θ , then for any estimator sequence Z_n of τ , one has

$$\lim_{b \rightarrow \infty} \lim_{c \rightarrow \infty} \lim_{n \rightarrow \infty} \inf_n \sup_{|\tau| \leq c} \mathbb{E}_{\theta_0 + \delta_n \tau} [\min(b, W(Z_n - \tau))] \geq \mathbb{E}[W(\xi)]$$

where ξ has a Gaussian distribution $\mathcal{N}(0, K^{-1})$. The lower bound is achieved by $Z_n = \delta_n^{-1}(T_n - \theta_0)$.

A loss function is “bowl-shaped” if the sublevel sets $\{u : W(u) \leq a\}$ are convex and symmetric around the origin. The value b is used to construct a bounded function $\min(b, W(Z_n - \tau))$. We let c go to infinity in order to cover a general case. The expectation $\mathbb{E}_{\theta_0 + \delta_n \tau}[\cdot]$ is taken w.r.t. a measure \mathcal{M} of the set $\{\theta : |\theta - \theta_0| \leq \delta_n \tau_n\}$ while $\mathbb{E}[\cdot]$ is taken w.r.t. a distribution of $K^{-1/2} \times \mathcal{N}(0, I)$ on ξ .

The theorem can be interpreted as follows. When using the auxiliary estimator θ_n^* in the likelihood ratio, this induces randomness to the local parameter τ_n . By using the LAN result in Theorem 3, we can attach the local parameter τ_n with a Gaussian measure. By the Gaussian prior assumption of τ_n , one can express the convergent procedure as a procedure of updating a Gaussian prior, while for a centered Gaussian prior, this procedure is to update the prior covariance matrix Γ^{-1} . The δ_n -regularity condition implies that K_n will converge uniformly in a neighborhood of θ_0 for arbitrary measure \mathcal{M} . Thus the covariance will converge to a the posterior covariance matrix $(K + \Gamma)^{-1}$. The Gaussian randomness introduces a new random variable ξ that has the posterior covariance matrix $(K + \Gamma)^{-1}$. The lower bound of the Bayes risk of this Gaussian variable is obtained by letting Γ go to zero, corresponding to initial values of τ widely spread. This is the local asymptotic minimax theorem. It is based on the minimax criterion and gives a lower bound for the maximum risk over a small neighborhood of the parameter θ . Because the local EL can achieve this lower bound, it is an asymptotically optimal estimator.

5 Simulations

Throughout the paper, our concerns are the violations of the standard regularity conditions for the moment restriction functions and their derivative functions. In this section, we simulate two models whose moment conditions are contaminated by some outliers. We call these models contaminated models. The contamination in this experiment occurs at a

Table 1:

Local Iteration			
Iter. num.	λ	τ	Local Estimator
1	0.148899	0.128070	2.069817
2	0.134007	0.115263	2.058290
3	0.120464	0.103737	2.047917
\vdots	\vdots	\vdots	\vdots
12	0.053311	0.084166	1.992690
13	0.043537	0.000000	1.992690

certain probability no matter how large the sample size is. The simulations try to mimic the environment that few observations may violate the boundedness condition for $m(X, \theta)$ and such observations are not caused by the small number of samples. In other words, some large values of $m_i(\theta)$ are caused by some x_i s and these x_i s are systematically existing.

These features imply that the specification of the constraint $\mathbb{E}[m(X, \theta)] = 0$ is invalid for the whole sample, although the specification is valid for the uncontaminated sample. A completely misspecified model is not of our interest. In our experiment, the contamination level is controlled to a small value so that the model is not significantly misspecified. A consequence of the mildly misspecified constraint is that the moment-based estimators are biased.

The full description of the localized EL's implementation is given in the Appendix. From each iteration in the localization steps, the value of the local estimator is adjusted. Table 1 gives an example of the information used in the localization step, where λ is the Lagrangian multiplier and τ is the local parameter. The true value of the parameter is 2. Due the mildly mis-speciation of the moment restriction, λ_n does not reach 0 when the estimator converges to the true value. However, the local iteration of τ induce an almost unbiased estimate result with the maximum local likelihood.

Figure 6.5 gives two representative phenomena in the numerical experiment. When the simulation does not induce a peculiar optimal point of the log-likelihood, EL rather than local EL reach the peak of the empirical log-likelihood function. However, such a peak is for the contaminated sample which induces misspecified moment restrictions. This peak does not lead to the best solution. Another situation is for irregular log-likelihood shape. In this case the EL estimation does not give a local optimal answer, nor even report a correct

Table 2:

Linear Model: LS Auxiliary Estimator								
c%=0.5%, L=10 (case I)					c%=0.005%, L=10000 (case II)			
Method	Mean	Median	MSE	IQR	Mean	Median	MSE	IQR
LS	2.092564	2.092765	0.009234	0.033984	2.091643	2.094945	0.009237	0.036843
IV	2.012655	2.012436	0.008073	0.130901	2.008087	2.012734	0.009991	0.148564
EL	1.990643	2.011329	0.024535	0.128907	1.984457	2.008713	0.033921	0.146062
Local EL(LS)	2.045564	2.050873	0.005683	0.077868	2.048954	2.056338	0.006055	0.084951

Linear Model: IV Auxiliary Estimator								
c%=0.05%, L=100 (case III)					c%=0.01%, L=10000 (case IV)			
Method	Mean	Median	MSE	IQR	Mean	Median	MSE	IQR
LS	2.091583	2.092988	0.009233	0.038327	2.091021	2.091687	0.009196	0.040935
IV	2.001433	2.003463	0.009768	0.132206	2.008599	2.001776	0.010176	0.137512
EL	1.985619	2.002222	0.028999	0.135985	1.983376	2.000420	0.032342	0.141040
Local EL(IV)	2.011766	2.015890	0.008267	0.120573	2.026002	2.021211	0.006859	0.105591

log-likelihood value. The problem is caused by the irregular shape of the likelihood. The flat log-likelihood region and the non-smooth peak break down the global search routine in the EL estimation. Although local EL estimate value does not correspond to the parameter value that gives the optimal log-likelihood for uncontaminated sample, local EL estimator reaches the local optimal point of the empirical log-likelihood function.

5.1 A Linear Experiment

We consider a simple structural model with a $n \times 1$ explanatory vector $\mathbf{x}_n = (x_1, \dots, x_n)^T$, a $n \times 1$ instrument vector \mathbf{z}_n and a disturbance vector \mathbf{u}_n , $n = 1000$. The parameter θ is equal to 2. The $n \times 1$ random vector ε is assumed to be normal. The model is as follows:

$$y_i = x_i\theta + \varepsilon_i,$$

$$x_i = z_i\pi + u_i.$$

In our numerical experiment, π is set to one. The instrument \mathbf{z}_n is a design vector with a constant vector plus a small noise and \mathbf{z}_n is independent of ε and \mathbf{u}_n . The uncertainty vector \mathbf{u}_n is a mixture of two normally distributed vector $\mathbf{u}_n^{(1)}$ and $\mathbf{u}_n^{(2)}$ where $u_i^{(1)} \sim \mathcal{N}(0, 1)$ and $u_i^{(2)} \sim \mathcal{N}(L, 1)$. L is referred to the degree of contamination. We introduce $u_i^{(2)}$ to generate

a mis-specified moment. In this experiment, $\mathbf{u}_n^{(2)}$ is a contaminated element. The mixing rate of $\mathbf{u}_n^{(2)}$ in \mathbf{u}_n is the probability of contamination. Let c denote this probability. If $P_{\mathbf{u}_n^{(i)}}$ denotes the distribution of $\mathbf{u}_n^{(i)}$, then $P_{\mathbf{u}_n} = (1 - c)P_{\mathbf{u}_n^{(1)}} + cP_{\mathbf{u}_n^{(2)}}$. We impose the correlation between ε and \mathbf{u}_n by using the equation $\varepsilon_n = R \times \mathbf{u}_n + \varepsilon'_n$ where $\varepsilon'_n \sim \mathcal{N}(0, 1)$ is independent of \mathbf{u}_n . The covariance value R is set to 0.1.

The moment restriction function in this example is $\mathbf{z}_n^T(\mathbf{y}_n - \mathbf{x}_n\theta)$. We will consider four different estimation methods, Least Squares (LS), Instrumental Variables (IV), EL, and local EL.¹¹ The estimators for LS, IV, EL are respectively $(\mathbf{x}_n^T \mathbf{x}_n)^{-1} \mathbf{x}_n^T \mathbf{y}_n$, $(\mathbf{z}_n^T \mathbf{x}_n)^{-1} \mathbf{z}_n^T \mathbf{y}_n$ and

$$\min_{\beta} \max_{\lambda_n} \sum_{i=1}^n \log(1 + \lambda_n m_i(\theta)),$$

where $m_i(\theta) = z_i(y_i - x_i\theta)$.

The true value of θ is 2. A consequence of the mildly misspecified constraint is that the moment-based estimators, IV and EL, are also biased but not as serious as LS. The bias of LS is caused by the correlation between ε and \mathbf{u}_n . Due to the endogenous problem, LS is always biased. The mild misspecified moment restriction leads to the small biases in IV and EL. We will use LS or EL as the auxiliary estimator of the local method. In Table 2, we show the estimation results for four cases: contamination percentage 0.5% (0.005%) with 10 (10000) degree of contamination with LS as an auxiliary estimator; contamination level 0.01% with 10 and 10000 degree of contamination with IV as an auxiliary estimator. The mean and the median of LS, IV and EL coincide with our expectation: a large bias in LS; a relative small bias in IV and EL. The level of bias in local method lies in-between. If one uses LS as the auxiliary estimator, then the bias of the local method is slightly larger than the case of using IV as the auxiliary estimator. However, among the four estimators, local EL attains the lowest mean square error (MSE) in all four cases. From the Q-Q plots in Figure 6.1 and 6.2, local EL is closer to the normal shape than EL. The density plots in Figure 6.3 and 6.4 show that the distribution of local EL is more concentrated in case (I) and (II) but its mean location is closer to the true value in case (III) and (IV).

¹¹In this setup, the IV estimator asymptotically has a degenerated second moment. Thus in order to make a fair comparison, we only consider the cases where the IV estimators are not widely spreaded.

5.2 A Nonlinear Experiment

We construct the moment restriction for a short-term interest rate model. Chan et al. (1992) show that the model can be nested within the following equations:

$$\begin{aligned} r_{t+1} - r_t &= \alpha + \beta r_t + \varepsilon_{t+1}, \\ \varepsilon_{t+1} &= \sigma r_t^\gamma u_t, \end{aligned}$$

where u_t is a normal white noise with zero mean and unit variance. α , β , γ , and σ are the parameters of the model. In this experiment, the contamination is introduced so that the distribution of u_t , P_{u_t} , is a mixture such that $(1 - c)P_{u_t^{(1)}} + cP_{u_t^{(2)}}$. As in the linear case, $u_t^{(1)} \sim \mathcal{N}(0, 1)$, $u_t^{(2)} \sim \mathcal{N}(L, 1)$ and c denotes the contaminated percentage. to a small value so that the model is not significantly misspecified.

Since we have four parameters, we construct the following four moments:

$$m_t(\theta) = \begin{bmatrix} \varepsilon_{t+1} \\ \varepsilon_{t+1}^2 - \sigma^2 r_t^{2\gamma} \end{bmatrix} \otimes \begin{bmatrix} 1 \\ r_t \end{bmatrix} = \begin{bmatrix} \varepsilon_{t+1} \\ \varepsilon_{t+1} r_t \\ \varepsilon_{t+1}^2 - \sigma^2 r_t^{2\gamma} \\ (\varepsilon_{t+1}^2 - \sigma^2 r_t^{2\gamma}) r_t \end{bmatrix},$$

where $\mathbb{E}[m_t(\theta)] = 0$. The sample moment restrictions are

$$\frac{1}{T} \begin{bmatrix} \sum_{t=1}^T ((r_{t+1} - \alpha - \beta r_t)) \\ \sum_{t=1}^T ((r_{t+1} - \alpha - \beta r_t) r_t) \\ \sum_{t=1}^T ((r_{t+1} - \alpha - \beta r_t)^2 - \sigma^2 r_t^{2\gamma} \Delta t) \\ \sum_{t=1}^T ((r_{t+1} - \alpha - \beta r_t)^2 - \sigma^2 r_t^{2\gamma} \Delta t) r_t \end{bmatrix}.$$

A consequence of the mildly misspecified constraint is that both GMM and EL are slightly biased. The biasness is caused by the contaminated u_t . Thus the auxiliary estimators of our local method are biased. In this model, we will only use EL as the auxiliary estimator.

We restrict the contaminated level to the moderate level by setting $L = 1000$. In the experiment, we select c to be 0.001% and 0.1%. Table 3 shows that the local result again lies in-between the alternative global results. In both cases, local EL reduces root of MSE of EL. But in the small contamination case (I), local EL is not as good as GMM because GMM over-performs EL. While in case (II), local EL becomes a better alternative. For estimates

Table 3:

Nonlinear Model						
	c%=0.001%, L=1000 (case I)			c%=0.1%, L=1000 (case II)		
Method	RMSE	IQR	MAD	RMSE	IQR	MAD
GMM	0.105864	0.053354	0.053786	2.613145	0.464843	4.027867
EL	0.106643	0.052433	0.053688	2.984457	0.509011	4.120982
Local EL	0.106532	0.052420	0.053711	2.607393	0.467948	4.087653

of each parameter, one can refer to the Q-Q plots in Figure 6.6 and 6.7.

6 Conclusion

We propose a new local EL method. We discuss its construction and derive theoretical properties. The construction is based on the infinite divisibility property; to the best of our knowledge, this feature has not yet been applied to EL. When the implied probability of EL is embedded in the infinitely divisible class, the log-likelihood ratio admits a local representation. Our local estimator is built on the basis of this representation. The consistency, local asymptotic normality, and asymptotic optimality of this estimator have been established. We apply the estimate method to two simulated experiments that require weaker regularity conditions for the estimator. The simulation results show that the local method reduces MSE from its auxiliary estimators.

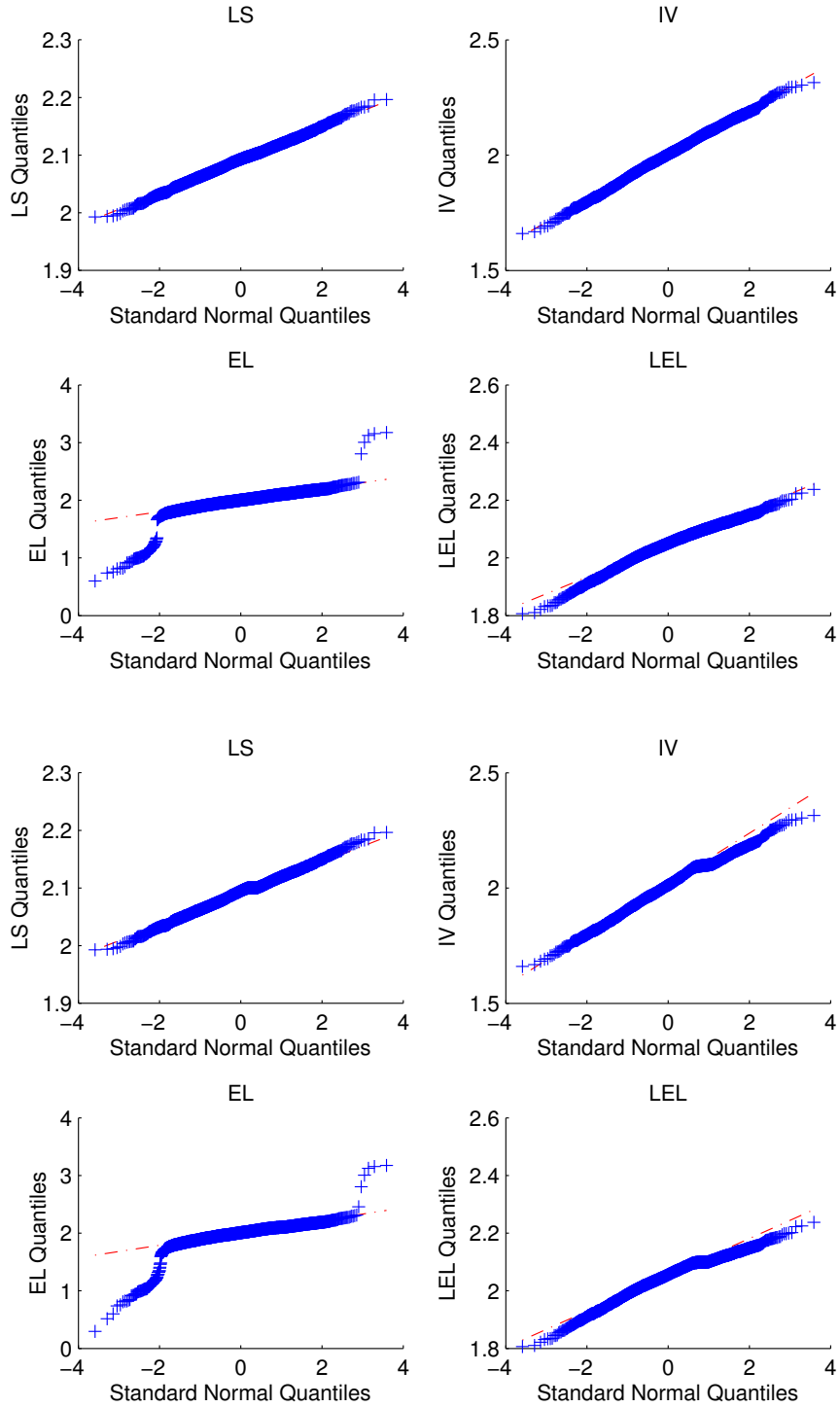


Figure 6.1: QQ plot (Densities of estimators). Up: case (I). Down: case (II).

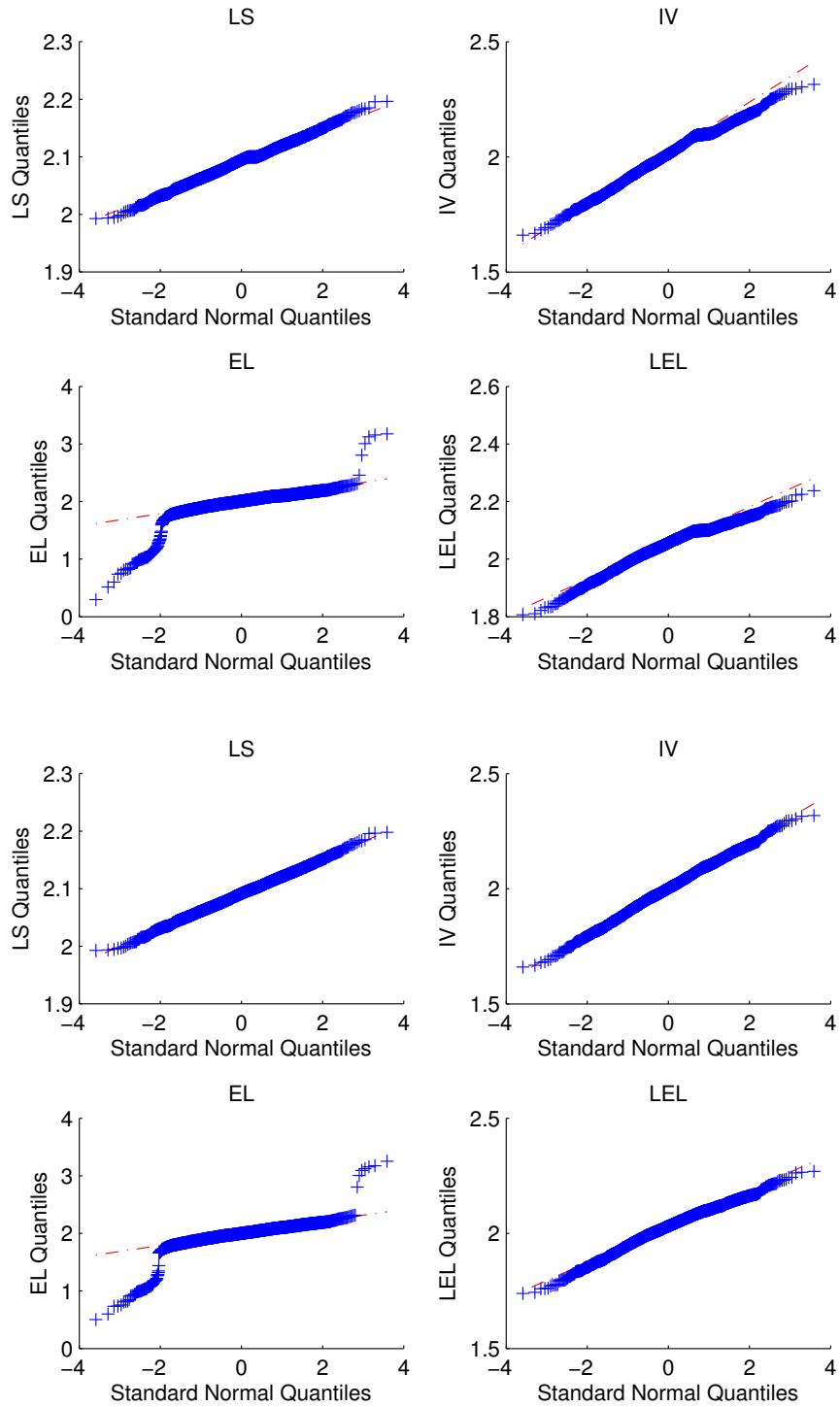


Figure 6.2: QQ plot (Densities of estimators). Up: case (III). Down: case (IV).

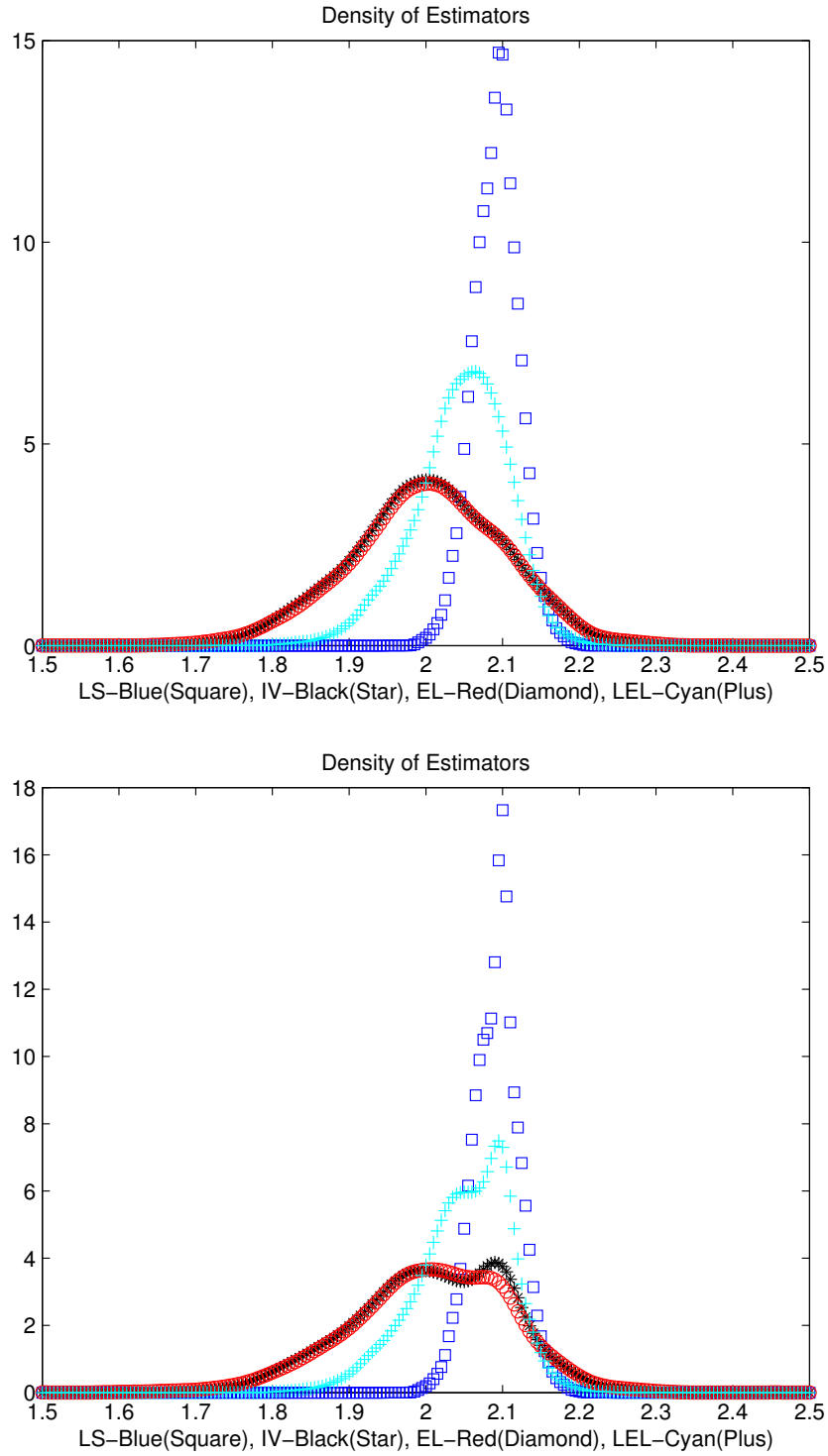


Figure 6.3: Density plot of four estimators: LS (Blue square), IV (Black star), EL (Red diamond) LEL (Cyan plus). Up: case (I). Down: case (II).

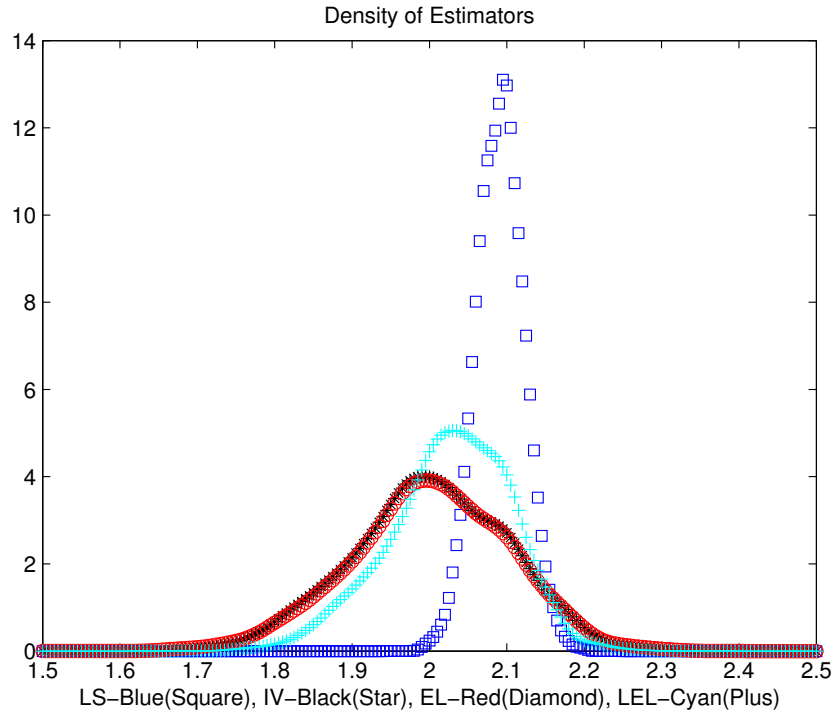
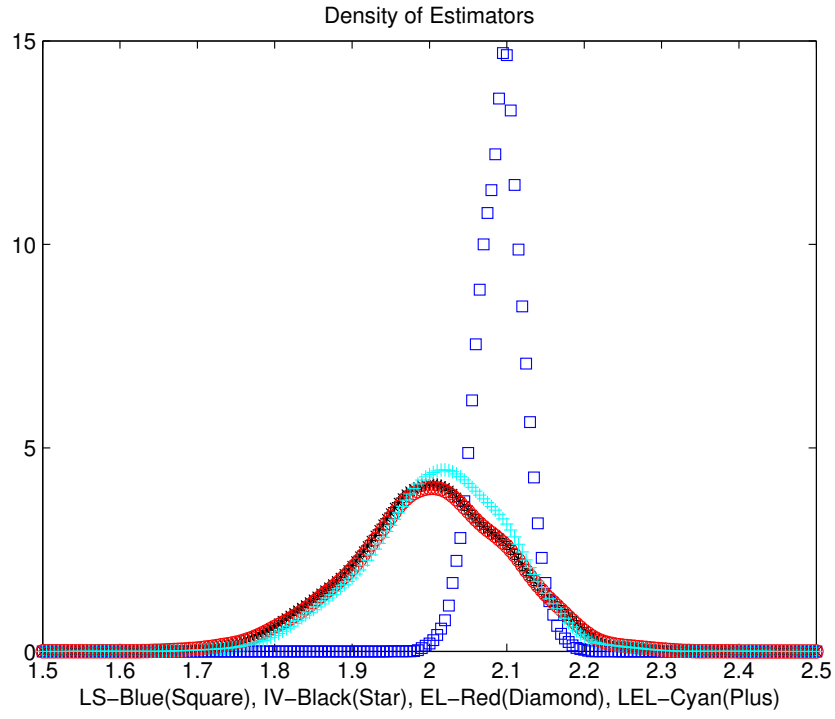


Figure 6.4: Density plot of four estimators: LS (Blue square), IV (Black star), EL (Red diamond) LEL (Cyan plus). Up: case (III). Down: case (IV).

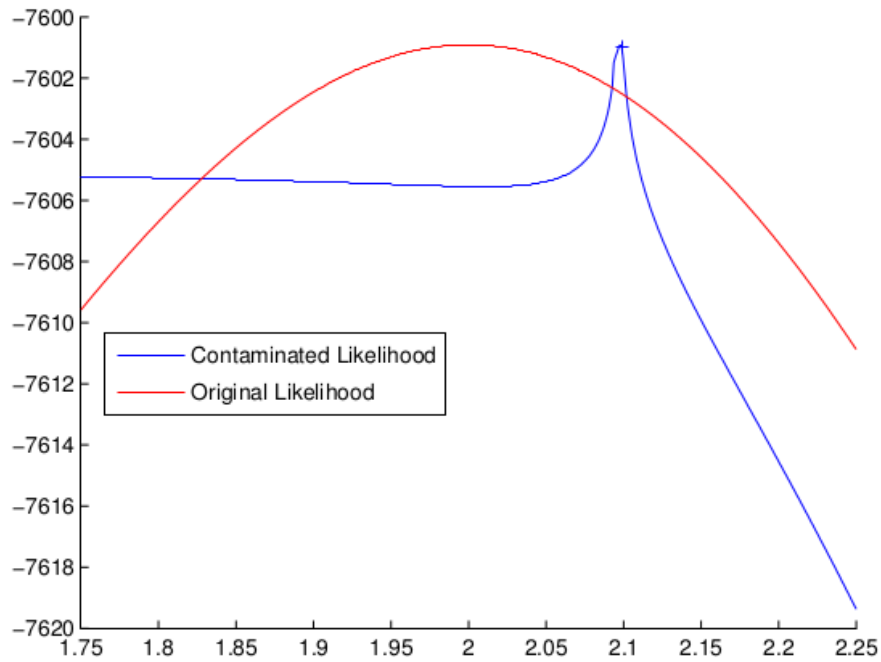
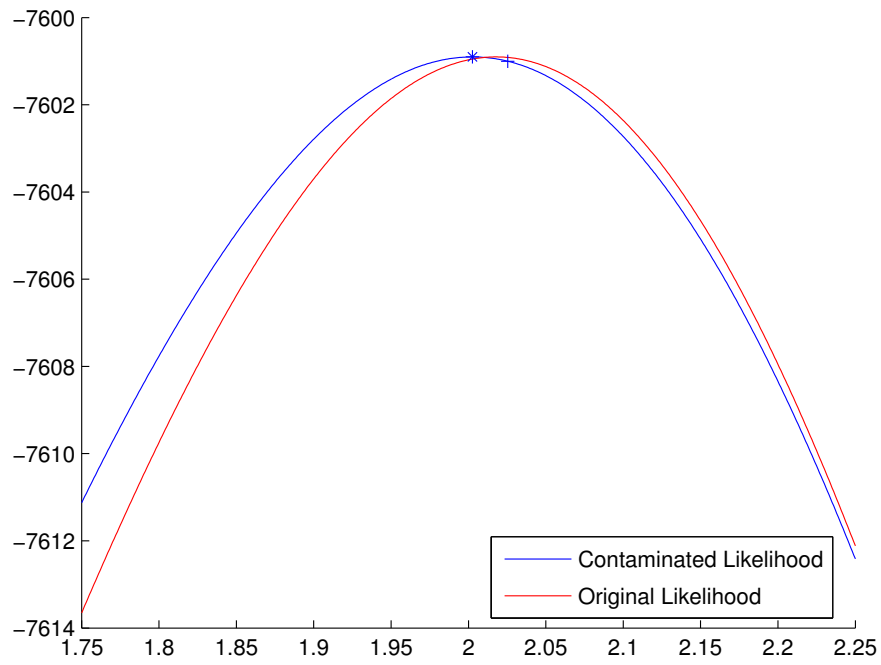


Figure 6.5: Log-likelihood. Cross stands for the LEL estimation result and star stands for the EL estimation result. Blue (Red) line is the log-likelihood for a contaminated (uncontaminated) sample.

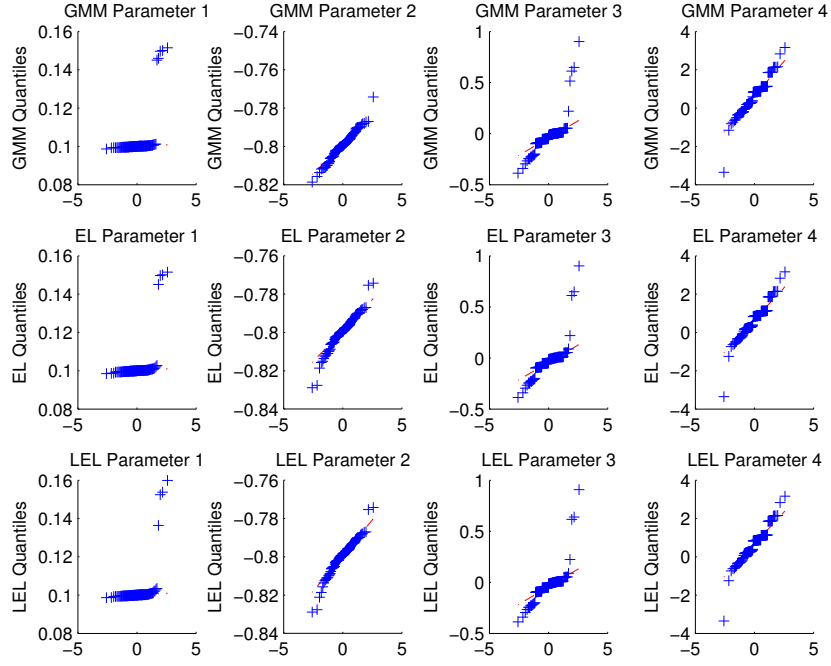


Figure 6.6: QQ plot (Densities of estimators). Case (I).

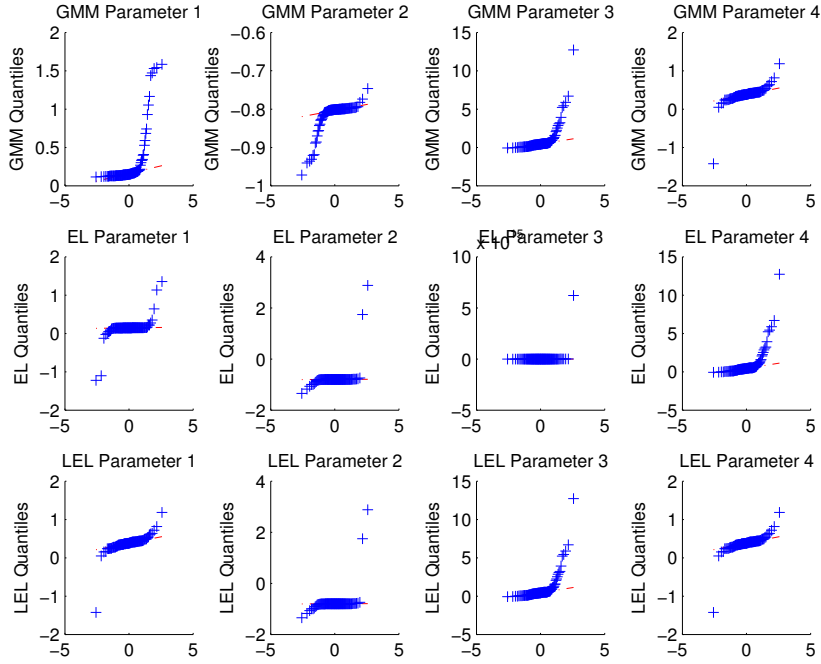


Figure 6.7: QQ plot (Densities of estimators). Case (II).

A Proof of Theorems

Proof of Theorem 1

The Lagrangian of EL is

$$L = \sum_{i=1}^n \log(np_i) - n\lambda^T \sum_{i=1}^n p_i m_i(\theta) - \gamma \left(\sum_{i=1}^n p_i - 1 \right),$$

where λ and γ are Lagrange multipliers. Setting the partial derivative of L w.r.t p_i equal to zero will give $\gamma = n$ and the implied probability $\tilde{p}_i = 1/(\gamma + n\lambda_n^T m_i(\theta))$. By the implicit function theorem, the partial derivative of $\sum_{i=1}^n \log \tilde{p}_i$ w.r.t λ gives a function $\Upsilon(\cdot, \cdot)$ of λ_n and θ such that

$$\begin{aligned} \frac{\partial \sum \log \tilde{\mathbf{p}}_i}{\partial \lambda} &:= \Upsilon(\lambda_n, \theta) = 0, \\ \implies \frac{1}{n} \sum_{i=1}^n \frac{m_i(\theta)}{1 + \lambda_n^T m_i(\theta)} &= \sum_{i=1}^n \tilde{p}_i(\theta) m_i(\theta) \end{aligned} \tag{A.1}$$

where λ_n is unique for fixed n and θ . Note that $\Upsilon(\lambda_n, \theta) = 0$ for $\forall \theta \in \Theta$ and θ is continuous hence $\Upsilon(\cdot)$ is continuous in θ . By the continuity of $m(X, \theta)$ and the representation of $\Upsilon(\cdot)$, we know that λ_n is also continuous on θ . The proof of the uniqueness of $\lambda(\theta)$ is as follows: because the set $\Gamma(\theta) = \lim_{n \rightarrow \infty} \cap_{i=1, \dots, n} \{\lambda | 1 + \lambda^T m(X_i, \theta) > 1/n\}$ is convex if it does not vanish, the function of $\log p$ is strictly concave on λ , so $\lambda(\theta)$ exists and is unique.

With these, the properties of likelihood ratio are shown in as follows. Equation (A.1) can be re-written as

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \left[1 - \frac{\lambda_n^T m_i(\theta)}{1 + \lambda_n^T m_i(\theta)} \right] m_i(\theta) &= 0 \\ \implies \frac{1}{n} \sum_{i=1}^n m_i(\theta) &= \frac{1}{n} \sum_{i=1}^n \frac{m_i(\theta) \lambda_n^T m_i(\theta)}{1 + \lambda_n^T m_i(\theta)} \\ &= \underbrace{\left[\sum_{i=1}^n \tilde{p}_i(\theta) m_i(\theta) m_i(\theta)^T \right]}_{(*)} \lambda_n. \end{aligned}$$

Condition 1 (v) states that $n^{-1} \sum_i m_i(\theta) m_i(\theta)^T$ is positive definite, let \mathbf{c} be larger than any eigenvalue of $n^{-1} \sum_i m_i(\theta) m_i(\theta)^T$ and let v be the corresponding eigenvector. The convex

combination of $m_i(\theta)m_i(\theta)^T$ over $\{\tilde{p}_i(\theta)\}$ in $(*)$ is bounded by $v^T \mathbf{c}v$. Let $E_v = v^T \mathbf{c}v$. According to condition 1 (iv), $m_i(\theta)$ has an envelop function $b(\theta)$ such that $\liminf_{\theta} |m(\theta, X)|/b(\theta) \geq 1$, then

$$\lim_{n \rightarrow \infty} |\lambda_n|/b'(\theta) \geq 1$$

for any θ where $b'(\theta) = b(\theta)/E_v$.

Let's first prove the existence of $\Lambda(\theta)$:

$$\begin{aligned} & \lim_{n \rightarrow \infty} \int \log \frac{1}{n} \frac{n}{1 + \lambda(n, \theta)^T m(x, \theta)} dP(x) \\ &= \mathbb{E} \lim_{n \rightarrow \infty} \log \frac{1}{1 + \lambda(n, \theta)^T m(X, \theta)} = \mathbb{E} \log \frac{1}{1 + \lambda(\theta)^T m(X, \theta)} = \Lambda(\theta). \end{aligned} \quad (\text{A.2})$$

The first convergence is by the LLN and the second equation is obtained by the dominated convergence Theorem, since $[1 + \lambda(\theta)^T m(x, \theta)]^{-1}$ is bounded and $\lambda(\theta)$ exists.

Next we prove the continuity of $\Lambda(\theta)$. The envelop functions $b'(\theta)$ and $b(\theta)$ are integrable and continuous (Condition 1), $\lambda(\theta)^T m(X, \theta)$ is bounded by a continuous function. Thus $\Lambda(\theta)$ is continuous and is bounded by an envelop function $b''(\theta) = \max(b'(\theta), b(\theta))$ such that

$$\sup_{\theta} \|\Lambda_n(\theta) - \Lambda(\theta)\| / b''(\theta) < 1 \quad (\text{A.3})$$

Now prove the identifiability of EL estimation. Choose a compact set $\Theta_c \subset \Theta$ such that for given ϵ

$$\sup_{\theta \in \Theta_c} |\Lambda(\theta)|/b''(\theta) \geq 1 - \epsilon.$$

By (A.3), LLN applied to $\Lambda_n(\theta)$ implies

$$\begin{aligned} \sup_{\theta} \frac{\|\Lambda_n(\theta) - \Lambda(\theta)\|}{b''(\theta)} &< \frac{\|\Lambda_n(\theta)\| - \|\Lambda(\theta)\| + 2\|\Lambda(\theta)\|}{b''(\theta)} - \epsilon \\ &< \frac{\|\Lambda_n(\theta)\| - \|\Lambda(\theta)\| + 2 \sup_{\theta \in \Theta_c} |\Lambda(\theta)|}{b''(\theta)} - \epsilon \\ &< \frac{\|\Lambda_n(\theta) - \Lambda(\theta)\| + 2 \sup_{\theta \in \Theta_c} |\Lambda(\theta)|}{b''(\theta)} < 1 - 3\epsilon \end{aligned}$$

The first inequality uses triangle inequality, the second one uses supremum property, and

the third one uses triangle inequality again. Therefore

$$\begin{aligned} |\Lambda_n(\theta) - \Lambda(\theta)| &\leq (1 - 3\epsilon)b''(\theta) \\ &\leq \frac{1 - 3\epsilon}{1 - \epsilon} \sup_{\theta \in \Theta_C} |\Lambda(\theta)| \leq (1 - \delta) \sup_{\theta \in \Theta_C} |\Lambda(\theta)| \end{aligned}$$

for $\forall \theta \in \Theta_c$. This inequality implies

$$\sup_{\theta \in \Theta_c} |\Lambda_n(\theta)| \leq \sup_{\theta \in \Theta_c} |\Lambda(\theta)| + \epsilon$$

asymptotically for any $\theta \in \Theta_c$. Thus if $\theta_0 \in \Theta_c$, then

$$\{T_n \subset \Theta_c\} \subset \left\{ \sup_{\theta \in \Theta_c} \Lambda_n(\theta) \leq \Lambda(\theta_0) + o_p(1) \right\},$$

where the probability of the event on the right side converges to one as $n \rightarrow \infty$. Because the compact set Θ could be shrinking to an arbitrary neighborhood of θ_0 , the EL estimator T_n is consistent.

Proof of Theorem 2

Before proving the theorem, we need to introduce a relation for univariate Gaussian families. For any pair of Gaussian measures in $\mathcal{G}_\Theta = \{G_\theta, \theta \in \Theta\}$, $G_\theta \subset \mathcal{E}_\theta$, there will be an expression to relate both of them as follows:

$$dG_\theta = \exp \left[\langle Y_\vartheta, \theta \rangle - \frac{1}{2} \|\theta\|^2 \right] dG_\vartheta, \quad (\text{A.4})$$

where $\vartheta, \theta \in \Theta$. The bilinear product in this expression is $\langle Y_\vartheta, \theta \rangle = \int_0^1 Y_\vartheta(t) G_\theta(dt)$ where Y_ϑ is a univariate Gaussian process. This is a random variable (functional integral or Wiener integral) with mean zero and variance $\|\theta\|^2 \leq \infty$ ¹². If dG_θ and dG_ϑ are defined as (A.4), the integral of $(dG_\theta/dG_\vartheta)^{1/2}$ w.r.t G_ϑ will has a linear quadratic representation.

Proof. Le Cam and Yang (2000, Proposition 4.1) show that the affinity between two Poissonized $d\tilde{P}_\theta, d\tilde{P}_\vartheta$ is

$$\int \sqrt{d\tilde{P}_\theta d\tilde{P}_\vartheta} = \exp \left\{ -\frac{1}{2} \|\theta - \vartheta\|^2 \right\}.$$

¹²This expression is called weak form expression and is often used for generalizing Gaussian processes.

Since Gnedenko and Kolmogorov (1968, Theorem 17.5) show that finite many number of Poisson type measures can approximate any infinitely divisible family and EL is embedded in an infinitely divisible family, we know the above expression is applicable over here. The Hellinger affinity for Gaussian family is

$$\int \sqrt{dG_\theta dG_\vartheta} = \int \exp \left[\frac{1}{2} \langle Y_\vartheta, \vartheta + \theta \rangle - \frac{1}{4} (\|\theta\|^2 + \|\vartheta\|^2) \right] dG_\vartheta.$$

The Gaussian property of $\langle Y_\vartheta, \vartheta + \theta \rangle$ implies that $\exp \left[\frac{1}{2} \langle Y_\vartheta, \vartheta + \theta \rangle \right]$ is log-normal distributed, then by log-normal property there is:

$$\int \exp \left[\frac{1}{2} \langle Y_\vartheta, \vartheta + \theta \rangle \right] dG_t = \exp \left(\frac{1}{8} \|\theta + \vartheta\|^2 \right).$$

Because only metric distance is going to be studied in $\int \sqrt{dG_\theta dG_\vartheta}$, we attach a Hilbert space to \mathcal{G} . The parallelogram identity for Hilbert space induces

$$\|\theta + \vartheta\|^2 + \|\theta - \vartheta\|^2 = 2 (\|\theta\|^2 + \|\vartheta\|^2),$$

so

$$2 (\|\theta\|^2 + \|\vartheta\|^2) + \|\theta + \vartheta\|^2 = -\|\vartheta - \theta\|^2$$

Therefore, $\sqrt{dG_\theta dG_\vartheta} = \exp(-\|\theta - \vartheta\|^2/8)$ is isometric to $\int \sqrt{d\tilde{P}_\vartheta d\tilde{P}_\theta} = \exp(-\|\theta - \vartheta\|^2/2)$.

If Fubini's theorem holds, the expression

$$2 \log \int \left(\frac{d\tilde{P}_\theta}{d\tilde{P}_\vartheta} \right)^{\frac{1}{2}} d\tilde{P}_\vartheta \approx 8 \log \int \left(\frac{dG_\theta}{dG_\vartheta} \right)^{\frac{1}{2}} dG_\vartheta$$

implies

$$\int \left[\log \frac{d\tilde{P}_\theta}{d\tilde{P}_\vartheta} \right] d\tilde{P}_\vartheta = 4 \int \left[\log \frac{dG_\theta}{dG_\vartheta} \right] dG_\vartheta$$

so that we can use the Gaussian expression (A.4) for the log-likelihood ratio process.

By Karhunen–Loeve Theorem (Kallenberg, 2002), the Gaussian process Y_θ can be expressed as

$$Y_\theta = \sum_{j=1}^{\infty} \xi_j \mathbf{u}_j(\theta)$$

where $\{\mathbf{u}_j\}$ constitutes an orthonormal basis for the Hilbert space \mathcal{G} and ξ_j are Gaussian

random variables and stochastically independent. Now let $\mathbf{u}_j(\cdot) = \sum_i^m \tau_i \mathbf{e}_i(\cdot)$ where \mathbf{e} is a unit basis for the local parameter space and τ_i are linear coefficients for $\mathbf{e}_i(\cdot)$. Let j indicate the index of a basis on the Hilbert space and i indicate the index of a basis on the local parameter space. Then the inner product in the Hilbert space can be expressed using local parameter coordinates such that $\langle Y_\vartheta, \theta \rangle = \sum_i^m \tau_i \theta_i \langle \mathbf{e}(\vartheta), \xi \rangle = \tau^T (\theta \tilde{\xi})$ where $\tilde{\xi}$ is also Gaussian because of the linear property. Let $\theta \tilde{\xi} = S'_\theta$ and $\mathbb{E}(\theta \tilde{\xi})^2 = K'_\theta$, then

$$\|\theta\|^2 = \mathbb{E}[\tau^T (\theta \tilde{\xi})]^2 = \tau^T K'_\theta \tau.$$

From (A.4), we have

$$\int \left[\log \frac{d\tilde{P}_\vartheta}{d\tilde{P}_\theta} \right] d\tilde{P}_\theta = \tau^T S_\theta - \frac{1}{2} \tau^T K_\theta \tau.$$

where $S_\theta = \int S'_\theta d\tilde{P}_\theta$ and $K_\theta = \int K'_\theta d\tilde{P}_\theta$. For a finite dimensional Gaussian vector based on n realizations Gaussian process, we have the sample counterparts τ_n , $S_{\theta,n}$ and $K_{\theta,n}$. We conclude that the EL ratio is approximately equal to the log-likelihood ratio of \mathcal{G} , which for the sample of size n is $\tau_n^T S_{\theta,n} - \tau_n^T K_{\theta,n} \tau_n / 2$. \square

Proof of Theorem 3

Proof. (i) When θ is given, by equation (3.1)

$$\begin{aligned} \Lambda_n(\theta + \delta_n \tau_n, \theta) &= \tau_n^T S_{\theta,n} - \frac{1}{2} \tau_n^T K_{\theta,n} \tau_n + o_{\tilde{p}_\theta}(1) \\ &= -\frac{1}{2} [(K_{\theta,n}^{-1} S_{\theta,n} - \tau_n^T)^T K_{\theta,n} (K_{\theta,n}^{-1} S_{\theta,n} - \tau_n^T)] \\ &\quad - (S_{\theta,n}^T K_{\theta,n}^{-1} S_{\theta,n}) + o_{\tilde{p}_\theta}(1). \end{aligned} \tag{A.5}$$

Similarly,

$$\Lambda_n(\theta + \delta_n \tau_n, \theta) = \tau_n^T K_n \delta_n^{-1} (T_n - \theta_n^*) - \frac{1}{2} \tau_n^T K_n \tau_n \tag{A.6}$$

$$\begin{aligned} &= -\frac{1}{2} [(\delta_n (T_n - \theta) - \tau_n^T)^T K_n (\delta_n (T_n - \theta) - \tau_n^T)] \\ &\quad - (\delta_n (T_n - \theta))^T K_n (\delta_n (T_n - \theta)). \end{aligned} \tag{A.7}$$

The difference between (A.5) and (A.6) tends to zero as $n \rightarrow \infty$. Non-negativity of K_n and $K_{\theta,n}$ shows that each of the four quadratic terms in (A.7) and A.5 must be non-negative. If

$S_{\theta,n}^T K_{\theta,n}^{-1} S_{\theta,n}$ converges to $(\delta_n(T_n - \theta))^T K_n (\delta_n(T_n - \theta))$, then

$$\begin{aligned} & (\delta_n(T_n - \theta) - \tau_n^T)^T K_n (\delta_n(T_n - \theta) - \tau_n^T) \rightarrow \\ & (K_{\theta,n}^{-1} S_{\theta,n} - \tau_n^T)^T K_{\theta,n} (K_{\theta,n}^{-1} S_{\theta,n} - \tau_n^T). \end{aligned}$$

So one can conclude that $K_n \rightarrow K_{\theta,n}$ and $S_n \rightarrow S_{\theta,n}$.

Now consider the opposite case $(\delta_n(T_n - \theta))^T K_n (\delta_n(T_n - \theta)) \not\rightarrow S_{\theta,n}^T K_{\theta,n}^{-1} S_{\theta,n}$. By a standard property of quadratic functions, we can have for some positive-definite matrix C

$$(\delta_n(T_n - \theta))^T K_n (\delta_n(T_n - \theta)) + C \rightarrow S_{\theta,n}^T K_{\theta,n}^{-1} S_{\theta,n}$$

Then for some vector Δ such that $\delta_n \Delta^T K_n \Delta \delta_n = C$, there is

$$(\delta_n(T_n - \theta + \Delta))^T K_n (\delta_n(T_n - \theta + \Delta)) \rightarrow S_{\theta,n}^T K_{\theta,n}^{-1} S_{\theta,n}$$

So $T_n + \Delta$ is optimal estimator for τ_n , because

$$\begin{aligned} & (\delta_n(T_n - \theta + \Delta) - \tau_n^T)^T K_n (\delta_n(T_n - \theta + \Delta) - \tau_n^T) \rightarrow \\ & (K_{\theta,n}^{-1} S_{\theta,n} - \tau_n^T)^T K_{\theta,n} (K_{\theta,n}^{-1} S_{\theta,n} - \tau_n^T). \end{aligned}$$

But this contradicts with our definition of T_n .

Thus $(\delta_n(T_n - \theta))^T K_n (\delta_n(T_n - \theta))$ converges to $S_{\theta,n}^T K_{\theta,n}^{-1} S_{\theta,n}$. It implies K_n converges to $K_{\theta,n}$ in probability and $\delta_n(T_n - \theta)$ converges to $K_{\theta,n}^{-1} S_{\theta,n}$.

(ii) By Proposition 2, we know that clustering points K_θ of $K_{\theta,n}$ are invertible. Since $\delta_n(T_n - \theta)$ converges to $K_{\theta,n}^{-1} S_{\theta,n}$, the limit of $\delta_n(T_n - \theta)$ is $K_\theta^{-1} S_{\theta,n}$. The Gaussian variable $S_{\theta,n}$ is second moment bounded. So the term $\delta_n(T_n - \theta)$ is bounded in probability.

(iii) We know the DQM condition implies (3.1), thus the linear-quadratic equation (3.1) may coincide with S_n and K_n by (i). The log-likelihood process can be rewritten as a centered log-likelihood process $\Xi_n(\cdot)$ plus a shift item $b_n(\cdot)$:

$$\delta_n \Lambda_n(\theta, \vartheta)(x) = \frac{1}{n} \delta_n \sum_{i=1}^n \log \frac{\tilde{p}_\theta}{\tilde{p}_\vartheta}(x_i) - \overbrace{\int \log \frac{\tilde{p}_\theta}{\tilde{p}_\vartheta}(x) dP_0}^{\Xi_n(\theta)}$$

$$+ \underbrace{\int \log \frac{\tilde{p}_\theta}{\tilde{p}_\vartheta}(x) dP_0}_{b_n(\theta)} + o_p(1).$$

Let $\delta_n = n^{-1/2}$. Given fixed $\lambda(\cdot)$ values in the constraint of equation (2.1), Theorem 2 says that $\log \frac{\tilde{p}_\theta}{\tilde{p}_\vartheta}(x_i)$ in $\Xi_n(\eta)$ can be replaced by a linear quadratic formulae w.r.t. τ_n , namely $\log \frac{\tilde{p}_\theta}{\tilde{p}_\vartheta}(x_i)$ belongs to a smooth functional class \mathcal{C}^2 . Therefore the process $\theta \mapsto \Xi_n(\theta)$ is an empirical process and $\Xi_n(\theta) \rightsquigarrow \Xi(\theta)$ by Donsker's Theorem, see van der Vaart (1998, Example 19.9) where $\Xi(\theta)$ is a Gaussian process. Note that $\Xi(\theta)$ has mean $\int \Xi(\theta) dP_0 = 0$ and covariance kernel $\mathbb{E}\Xi^2(\theta)$ under P_0 . The log-normal property implies that $\mathbb{E} \exp[\Xi(\theta) + b(\theta)] = 1$ with the expectation taken under P_0 and $b(\theta) = \lim_{n \rightarrow \infty} b_n(\theta)$. Log normal property of $\exp \Xi(\cdot)$ gives $b(\theta) = -(1/2)\mathbb{E}\Xi^2(\theta)$. By Proposition 1 and equation (3.1), we can show that

$$\begin{aligned}\Xi_n(\theta) &= S_{\theta,n} \\ b(\theta) &= -\frac{1}{2}K_\theta,\end{aligned}$$

and when $\theta = \theta_0$

$$\begin{aligned}\Xi_n(\theta_0) &= \mathbb{E} \left[\frac{\partial m(X, \theta_0)^T}{\partial \theta} (\mathbb{E} m(X, \theta_0) m(X, \theta_0)^T)^{-1} \right] \delta_n \sum_i^n m_i(\theta_0) \\ b(\theta_0) &= -\frac{1}{2} \mathbb{E} \frac{\partial m(X, \theta_0)^T}{\partial \theta} (\mathbb{E} m(X, \theta_0) m(X, \theta_0)^T)^{-1} \mathbb{E} \frac{\partial m(X, \theta_0)}{\partial \theta}.\end{aligned}$$

□

Proof of Theorem 4

Discussion: The proof follows the strategies of van der Vaart (Proposition 8.6 1998) and Le Cam and Yang (Theorem 6.1 1990). The difficulty comes from the expectation conditional on the local parameter τ . Note that the measure \mathcal{M} has not yet been specified. If one can in Bayesian fashion give a prior distribution on \mathcal{M} , then what we need to study is the posterior distributions given this “local prior measures”. In fact, the δ_n -sparse condition already implies that for arbitrary priors, the corresponding posteriors concentrate on the small shrinking neighborhood of θ_0 .

Proof. First look at the population log-likelihood ratio

$$\Lambda(\theta + \tau, \theta) = -\frac{1}{2} \left[(K_\theta^{-1} S_\theta - \tau)^T K_\theta (K_\theta^{-1} S_\theta - \tau) - (S_\theta^T K_\theta^{-1} S_\theta) \right] + o_{\tilde{P}_\theta}(1).$$

which implies that the term $(K_\theta^{-1} S_\theta - \tau)^T K_\theta (K_\theta^{-1} S_\theta - \tau)$ is χ^2 distributed. The quadratic form of a Gaussian variable ξ , $\xi^T \xi$, can generate exactly the same distribution. As Theorem 2 shows that the approximation of Gaussian family is feasible. For any value of θ , there will be such a ξ_θ whose distribution is equivalent to $K_\theta^{-1} S_\theta - \tau$ and has the variance $K_\theta^{-1/2}$. Then we have the expression

$$\tau = K_\theta^{-1} S_\theta - \xi_\theta,$$

which shows that τ consists of two Gaussian variables $K_\theta^{-1} S_\theta$ and ξ_θ . Thus we are able to impose a Gaussian structure on the measure \mathcal{M} .

Now we can look at the expectation $\min(b, \mathbb{E}[W(Z_n - \tau)|\theta_0 + \delta_n \tau])$ which is bounded by b . Since both “prior” and “posterior” concentrate around θ_0 and are Gaussian, the updating information only occurs for covariance matrix. Let τ be a Gaussian random variable centered at 0 with inverse covariance Γ . The conjugate property indicates the posterior of τ can be written as:

$$Z_n = \delta_n^{-1}(\tilde{T}_n - \theta_0) = (K_n + \Gamma)^{-1/2} K_n \delta_n^{-1} (T_n - \theta_0),$$

especially when $\Gamma = 0$, $Z_n = \delta_n^{-1}(T_n - \theta_0)$. By Anderson’s Lemma¹³, for bounded W , there is

$$\mathbb{E}[W(Z_n - \tau)|\theta_0 + \delta_n \tau] \geq \mathbb{E}[W(Z_n)|\theta_0 + \delta_n \tau].$$

Since $K_n \delta_n^{-1}(T_n - \theta_0) \sim \mathcal{N}(0, I)$, the lower bound of $\mathbb{E}[W(Z_n - \tau)|\theta_0 + \delta_n \tau]$ is

$$\mathbb{E} \left\{ W \left[(K_n + \Gamma)^{-1/2} \times \mathcal{N}(0, I) \right] \mid K_n + \Gamma \right\}.$$

The measure of $\theta_0 + \delta_n \tau$ is replaced by $K_n + \Gamma$ because of the Gaussian property, namely the update of covariance matrix. Note that K_n and Γ are independent with $\mathcal{N}(0, I)$. With the condition $K_n \rightsquigarrow K_\theta$ in \tilde{P}_θ law, the limit becomes $\mathbb{E} \left\{ W \left[(K_\theta + \Gamma)^{-1/2} \times \mathcal{N}(0, I) \right] \right\}.$

¹³For a symmetric distribution, shifting an integral function of it to a new position will product higher expected value, see van der Vaart (1998, Lemma 8.5).

When c is very large, the probability of normal prior $|\tau| > c$ is small enough thus

$$\liminf_n \sup_{|\tau| \leq c} \mathbb{E} \{ W [(K_n + \Gamma)^{-1/2} \times \mathcal{N}(0, I)] \} \geq \mathbb{E} \{ W [(K_\theta + \Gamma)^{-1/2} \times \mathcal{N}(0, I)] \} - \Delta$$

for small enough Δ . Especially, when Γ go to zero or say the measure \mathcal{M} degenerates to a point eventually, $Z_n = \delta_n^{-1}(T_n - \theta_0)$ obtains the lower bound $\mathbb{E}[W(K_\theta^{-1/2}) \times \mathcal{N}(0, I)]$. If $W = 1$ and $K_\theta = K$, by Theorem 3(iii) we achieve the efficient bound of semi-parametric estimators. \square

B Other Technical Details

Poisson Approximation for Arbitrary Infinitely Divisible Families

Let $\phi(t)$ and $\phi_n(t)$ be the characteristic functions of distributions in \mathcal{E} and \mathcal{E}_n . By the infinitely divisible property, $\phi(t) = [\phi_n(t)]^n$ or $\phi_n(t) = [\phi(t)]^{1/n}$. Two characteristic functions have the following relation:

$$\begin{aligned} n(\phi_n(t) - 1) &= n(\sqrt[n]{\phi(t)} - 1) = n \left(e^{\frac{1}{n} \log \phi(t)} - 1 \right) \\ &= n \left(1 + \frac{1}{n} \log \phi(t) + o\left(\frac{1}{n}\right) - 1 \right) \rightarrow \log \phi(t), \end{aligned}$$

or say $\exp(n(\phi_n(t) - 1)) \rightarrow \phi(t)$. The concrete construction of characteristic function in $\mathcal{E}_{\theta,n}$ depends on the discrete Fourier transform of $\Lambda(X, \theta)$ on j segments e.g. $\inf \Lambda(X) < c_1 < c_2 < \dots < c_j < \sup \Lambda(X)$ which implies that

$$\lim_{j \rightarrow \infty} \sum_{k=1}^j a_k(i) e^{itc_k} = \int e^{it\Lambda(X)} dF_n = \phi_n(t),$$

where $a_n(k) = n(F_n(c_k) - F_n(c_{k-1}))$ is the Fourier coefficient¹⁴ and F_n is the measure for $\Lambda_n(\theta)$. Combined with the expression above, one can see that a characteristic function of

¹⁴The Stieltjes sum, a discrete version of stochastic integral.

finite many number of Poisson measures (compound Poisson measures) approximates $\phi(t)$:

$$\exp \sum_{i=1}^j (na_i) (e^{it\Lambda(x_i, \theta)} - 1) \rightarrow \phi(t) \quad (\text{B.1})$$

where $j \rightarrow \infty$ and $\{na_i\}_{i=1, \dots, j}$ converges to a measure. To see the argument of (B.1), let $V(\cdot)$ be a Poisson process (a random measure) with Poisson parameter γ such that $\mathbb{E}V(\mathcal{A}) = \gamma(\mathcal{A})$ for a set \mathcal{A} . For any function v in infinite divisible family, the characteristic function of v is $\phi(t) = \exp\{\int (e^{itv} - 1)d\gamma\}$.

The approximation can be viewed as constructing a new family which approximately equals the infinite divisible \mathcal{E}_θ . Firstly select a Poisson variable ν (again a random measure) such that $\mathbb{E}\nu(\Lambda(X)) = 1$ for any log-likelihood ratio $\Lambda(X)$ and then carry out n -draws from the direct product $\otimes_{i=1, \dots, \nu} \mathcal{E}_{\theta, i}$, ν copies $\mathcal{E}_{\theta, i}$. The result is called a poissonized family.

Derivation of Equation (3.3)

Since $\mathbb{E}[\exp(\log(dG_i/d\mu))] = 1$, then we have

$$\mathbb{E} \exp \left[L(i) + \mathbb{E} \log \left(\frac{dG_i}{d\mu} \right) \right] = [\mathbb{E} e^{L(i)}] \cdot e^{\mathbb{E} \log(\frac{dG_i}{d\mu})} = 1$$

By the log-normal property, $\mathbb{E} \exp L(i) = e^{\frac{1}{2}K(i, i)}$, we have

$$e^{\frac{1}{2}K(i, i)} \cdot e^{\mathbb{E} \log(\frac{dG_i}{d\mu})} = 1 \iff \mathbb{E} \left[\log \left(\frac{dG_i}{d\mu} \right) \right] = -\frac{1}{2}K(i, i)$$

thus we have (3.2). For $\mathbb{E} \exp[L(\theta) + L(\vartheta)]$, we have $2K(\theta, \vartheta)$. Combining $2K(\theta, \vartheta)$ and $K(i, i)$ gives us (3.3).

Proof of Proposition 1

The proof is based on Taylor expansions. Note that

$$m(x, \theta_0 + \delta_n \tau) = m(x, \theta_0) + \delta_n \frac{\partial m(x, \theta_0)}{\partial \theta^T} \tau + o_p(\delta_n^2). \quad (\text{B.2})$$

Let $\theta \in \{\theta \mid |\theta - \theta_0| \leq |\tau| \delta_n\}$, $|\tau|$ is a vector with elements equal to their absolute values. The result

$$\lambda_n(\theta) = \left(\sum_{i=1}^n [m_i(\theta) m_i(\theta)^T] / n \right)^{-1} \sum_{i=1}^n m_i(\theta) / n + o_p(n^{-1/2})$$

holds uniformly for θ in a neighborhood of θ_0 , see the proofs in Qin and Lawless (1994, Lemma 1) or Owen (2001, Theorem 2.2). For the empirical log-likelihood at θ , by noting that $\lambda_n^T m_i$ is close to zero and using a second order approximation for $\log(1 + \lambda_n^T m_i)$, we obtain:

$$\begin{aligned} \sum_{i=1}^n \log \tilde{p}_\theta &= \sum_{i=1}^n \left[\lambda_n(\theta)^T m_i(\theta) - \frac{1}{2} (\lambda_n(\theta)^T m_i(\theta) m_i(\theta)^T \lambda_n(\theta)) \right] \\ &\quad - n \log n + o_p(1). \end{aligned}$$

The remainder term is based on bounding $\sum_{i=1}^n (\lambda_n^T m_i)^3$ for which Owen (1990) showed in Lemma 3 that it is of order $o_p(1)$. Note that his γ_i is our $\lambda_n^T m_i(\theta)$. Note that

$$\lambda_n(\theta)^T m_i(\theta) = \left(\sum_{i=1}^n \frac{m_i(\theta)}{n} \right)^T \left[\sum_{i=1}^n \frac{1}{n} (m_i(\theta) m_i(\theta)^T) \right]^{-1} m_i(\theta)$$

and after summation equals the squared term:

$$\begin{aligned} &\sum_{i=1}^n \lambda(\theta)^T m_i(\theta) m_i(\theta)^T \lambda_n(\theta) = \\ &\left(\sum_{i=1}^n \frac{m_i(\theta)}{n} \right)^T \left[\sum_{i=1}^n \frac{1}{n} (m_i(\theta) m_i(\theta)^T) \right]^{-1} \left(\sum_{i=1}^n \frac{m_i(\theta)}{n} \right). \end{aligned}$$

So adding these two terms we obtain:

$$\begin{aligned} \sum_{i=1}^n \log \tilde{p}_\theta &= \frac{1}{2} \left(\sum_{i=1}^n \frac{m_i(\theta)}{n} \right)^T \left[\sum_{i=1}^n \frac{1}{n} (m_i(\theta) m_i(\theta)^T) \right]^{-1} \\ &\quad \times \left(\sum_{i=1}^n \frac{m_i(\theta)}{n} \right) - n \log n + o_p(1). \end{aligned}$$

It implies:

$$\begin{aligned}
2 \sum_{i=1}^n \log \frac{\tilde{p}_{\theta_0 + \delta_n \tau}(x_i)}{\hat{p}_{\theta_0}}(x_i) &= \left(\frac{1}{n} \sum_{i=1}^n m_i(\theta_0 + \delta_n \tau) \right)^T \times \\
&\quad \left(\frac{1}{n} \sum_{i=1}^n [m_i(\theta_0 + \delta_n \tau) m_i(\theta_0 + \delta_n \tau)^T] \right)^{-1} \sum_{i=1}^n m_i(\theta_0 + \delta_n \tau) - \\
&\quad \left(\frac{1}{n} \sum_{i=1}^n m_i(\theta_0) \right)^T \left(\frac{1}{n} \sum_{i=1}^n [m_i(\theta_0) m_i(\theta_0)^T] \right)^{-1} \sum_{i=1}^n m_i(\theta_0) + o_p(1).
\end{aligned}$$

It follows from the approximation of λ above. Using equation (B.2) we can further simplify the terms involving $\theta + \delta_n \tau$. We obtain for the middle term:

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n [m_i(\theta_0 + \delta_n \tau) m_i(\theta_0 + \delta_n \tau)^T] &= \frac{1}{n} \sum_{i=1}^n [m_i(\theta_0) m_i(\theta_0)^T] + \\
&\quad \delta_n \tau \left(\frac{\partial m_i(\theta_0)}{\partial \theta^T} \right)^T m_i(\theta_0) + \frac{(\delta_n \tau)^2}{4} \left(\frac{\partial m_i(\theta_0)}{\partial \theta^T} \right)^T \frac{\partial m_i(\theta_0)}{\partial \theta^T} + o_p(\delta_n^3) \\
&= \frac{1}{n} \sum_{i=1}^n [m_i(\theta_0) m_i(\theta_0)^T] + \frac{1}{n} \delta_n O_p(n^{1/2}) + o_p(\delta_n^2) + o_p(\delta_n^3).
\end{aligned}$$

With the big bracket becoming

$$\begin{aligned}
&n \left[\frac{1}{n} \sum_{i=1}^n m_i(\theta_0) + \frac{1}{n} \sum_{i=1}^n \delta_n \frac{\partial m_i(\theta_0)}{\partial \theta^T} \tau \right]^T \left(\frac{1}{n} \sum_{i=1}^n [m_i(\theta_0) m_i(\theta_0)^T] \right)^{-1} \\
&\quad \times \left[\frac{1}{n} \sum_{i=1}^n m_i(\theta_0) + \frac{1}{n} \sum_{i=1}^n \delta_n \frac{\partial m_i(\theta_0)}{\partial \theta^T} \tau \right] \\
&= n \left[\frac{1}{n} \sum_{i=1}^n m_i(\theta_0) + \delta_n \mathbb{E} \frac{\partial m_i(\theta_0)}{\partial \theta^T} \tau + \delta_n O(n^{-1/2} (\log \log n)^{1/2}) \right]^T \\
&\quad \times (\mathbb{E} (m(x, \theta_0) m(x, \theta_0)^T))^{-1} \\
&\quad \times \left[\frac{1}{n} \sum_{i=1}^n m_i(\theta_0) + \delta_n \mathbb{E} \frac{\partial m_i(\theta_0)}{\partial \theta^T} \tau + \delta_n O(n^{-1/2} (\log \log n)^{1/2}) \right] \\
&= 2 \delta_n \mathbb{E} \frac{\partial m(x, \theta_0)}{\partial \theta^T} \tau (\mathbb{E} (m(x, \theta_0) m(x, \theta_0)^T))^{-1} \frac{1}{n} \sum_{i=1}^n m_i(\theta_0) + \\
&\quad \delta_n^2 \mathbb{E} \frac{\partial m(x, \theta_0)}{\partial \theta^T} \tau (\mathbb{E} (m(x, \theta_0) m(x, \theta_0)^T))^{-1} \mathbb{E} \frac{\partial m(x, \theta_0)}{\partial \theta} \tau +
\end{aligned}$$

$$\frac{1}{n} \sum_{i=1}^n m_i(\theta_0) \left(\mathbb{E} (m(x, \theta_0) m(x, \theta_0)^T) \right)^{-1} \frac{1}{n} \sum_{i=1}^n m_i(\theta_0) + o_p(\delta_n^3)$$

where $O(n^{-1/2}(\log \log n)^{1/2})$ is used to bound the difference of the sample average and the expectation of a random vector. Thus the local EL is

$$2 \sum_{i=1}^n \log \frac{\tilde{p}_{\theta_0 + \delta_n \tau_n}(x_i)}{\tilde{p}_{\theta_0}} = \delta_n \tau_n^T A_1 + \frac{1}{2} \delta_n^2 \tau_n^T A_2 \tau_n + o_p(1)$$

where

$$A_1 = \mathbb{E} \frac{\partial m(X, \theta_0)^T}{\partial \theta} \left(\mathbb{E} m(X, \theta_0) m(X, \theta_0)^T \right)^{-1} \sum_{i=1}^n m_i(\theta_0),$$

$$A_2 = \mathbb{E} \frac{\partial m(X, \theta_0)^T}{\partial \theta} \left(\mathbb{E} m(X, \theta_0) m(X, \theta_0)^T \right)^{-1} \mathbb{E} \frac{\partial m(X, \theta_0)}{\partial \theta^T}.$$

Note that $O(n^{-1/2}(\log \log n)^{1/2}) \times \delta_n \sum_{i=1}^n m_i(\theta_0) = o_p(1)$ and

$$\lim_{n \rightarrow \infty} A_n \cdot \sum_{i=1}^n [m_i(\theta_0 + \delta_n \tau) - m_i(\theta_0)]/n = o_p(1)$$

with $A_n = \sum_{i=1}^n m_i(\theta_0) \left(\mathbb{E} m(x, \theta_0) m(x, \theta_0)^T \right)^{-1}$ by the continuity of $m_i(\theta)$.

Proof of Proposition 2

To prove K_θ is invertible, we will prove K_θ is almost surely positive definite. Le Cam's first Lemma implies that

$$\mathbb{E} \exp \left[\tau^T S_\theta - \frac{1}{2} \tau^T K_\theta \tau \right] = 1. \quad (\text{B.3})$$

Because (B.3) holds for all τ , we can use a symmetrized method to simplify (B.3). For a given value τ and $-\tau$, we have

$$\mathbb{E} \left\{ \exp \left[\tau^T S_\theta - \frac{1}{2} \tau^T K_\theta \tau \right] + \exp \left[-\tau^T S_\theta - \frac{1}{2} \tau^T K_\theta \tau \right] \right\} = 2.$$

By $\cosh \tau^T S_\theta = (\exp \tau^T S_\theta + \exp(-\tau^T S_\theta))/2$, we have

$$\mathbb{E}[(\cosh \tau^T S_\theta) \exp(-\tau^T K_\theta \tau / 2)] = 1. \quad (\text{B.4})$$

Assume there is some τ such that $\tau^T K_\theta \tau$ is negative, then

$$\begin{aligned} & \mathbb{E} [\mathbb{I}_{\{\tau^T K_\theta \tau > 0\}} (\cosh \tau^T S_\theta) \exp(-\tau^T K_\theta \tau / 2)] \\ & \leq \mathbb{E} [(\cosh \tau^T S_\theta) \exp(-\tau^T K_\theta \tau / 2)] = 1 \end{aligned} \tag{B.5}$$

where $\mathbb{I}_{\{\cdot\}}$ is an indicator function. However, since

$$\exp(-\tau^T K_\theta \tau / 2) > 1$$

when $\tau^T K_\theta \tau$ is negative and $(\cosh \tau^T S_\theta) > 1$,

$$\begin{aligned} & \underbrace{\mathbb{E} [\mathbb{I}_{\{\tau^T K_\theta \tau > 0\}} (\cosh \tau^T S_\theta) \exp(-\tau^T K_\theta \tau / 2)]}_{>0} + \\ & \underbrace{\mathbb{E} [\mathbb{I}_{\{\tau^T K_\theta \tau \leq 0\}} (\cosh \tau^T S_\theta) \exp(-\tau^T K_\theta \tau / 2)]}_{\geq 1} \\ & = \mathbb{E} [(\cosh \tau^T S_\theta) \exp(-\tau^T K_\theta \tau / 2)] > 1 \end{aligned}$$

we have a contradiction with equation (B.4) unless the set $\{\tau^T K_\theta \tau \leq 0\}$ is empty. Therefore, K_θ is positive definite and hence invertible.

C Implementation of the Local EL in Section 4

The evaluation of the LEL estimator requires evaluation of S_n and K_n . It appears reasonable to use any numerical first and the second derivative of $\Lambda_n(\theta_n^* + \delta_n \tau, \theta_n^*)$. The matrix $u_i^T K_n u_j = \{K_{n,i,j}\}$ in Section 4

$$\begin{aligned} K_{n,i,j} = & - \{ \Lambda_n[\theta_n^* + \delta_n(u_i + u_j), \theta_n^*] \\ & - \Lambda_n[\theta_n^* + \delta_n u_i, \theta_n^*] - \Lambda_n[\theta_n^* + \delta_n u_j, \theta_n^*] \} \end{aligned}$$

is a particular form of a numerical derivatives. If $u \in \mathbb{R}$, the above expression of $K_{n,i,j}$ can be simplified to

$$K_n = \frac{-\Lambda_n[\theta_n^* + \delta_n 2u, \theta_n^*] + 2\Lambda_n[\theta_n^* + \delta_n u, \theta_n^*]}{u^2}.$$

For a fixed value of δ_n , if we let $f(\delta_n u) = \Lambda_n[\theta_n^* + \delta_n 2u, \theta_n^*]$, then there is

$$\delta_n^{-2} K_n = - \frac{[f(2\delta_n u) - f(\delta_n u)]/\delta_n u - [f(\delta_n u) - f(0)]/\delta_n u}{\delta_n u}$$

which is a simple one-sided numerical second derivative of $f(\delta_n u)$ at $u = 0$, multiplied by -1 . Note that

$$\lim_{\delta_n u \rightarrow 0} \frac{\Lambda_n[\theta_n^* + \delta_n u, \theta_n^*] - 0}{\delta_n u}$$

define the derivative of $\Lambda_n[\theta_n^* + \delta_n u, \theta_n^*]$ at θ_n^* . In our implementation, instead of using the expression $(f(\delta_n u) - f(0))/\delta_n u$, we will focus on the derivative form of $\Lambda_n[\theta_n^* + \delta_n u, \theta_n^*]$. While $\lambda_n(\theta_n^*)$ in $\tilde{p}_{\theta_n^*}$ cannot be attained as a closed form expression, we will use the Romberg method to handle this difficulty.

The whole implementation of LEL follows the definition in Section 4.

Step 1. Find an auxiliary estimate θ_n^* using LS or IV.

Step 2. can be written as an expression of a 2nd order finite difference

$$\begin{aligned} K_{n,i,j} &= - \left\{ \log \frac{\tilde{p}_{\theta_n^* + 2\delta_n u}}{\tilde{p}_{\theta_n^*}} - \log \frac{\tilde{p}_{\theta_n^* + \delta_n u}}{\tilde{p}_{\theta_n^*}} - \log \frac{\tilde{p}_{\theta_n^* + \delta_n u}}{\tilde{p}_{\theta_n^*}} \right\} \\ &= - \left\{ \frac{1}{2} \cdot 2 (\log \tilde{p}_{\theta_n^* + 2\delta_n u} - \log \tilde{p}_{\theta_n^*}) - \right. \\ &\quad \left. 2 (\log \tilde{p}_{\theta_n^* + \delta_n u} - \log \tilde{p}_{\theta_n^*}) \right\} \\ &= -[f(2\delta_n u) - f(\delta_n u)] - [f(\delta_n u) - f(0)] \end{aligned}$$

Then $(f(\delta_n u) - f(0))/\delta_n u$ can be expressed as a directional derivative $\frac{\partial}{\partial \vec{u}} \log \tilde{p}$ evaluated at θ_n^* :

$$\frac{1}{\delta_n u} (\log \tilde{p}_{\theta_n^* + \delta_n u_i} - \log \tilde{p}_{\theta_n^*}) \rightarrow \frac{\partial}{\partial \vec{u}} \log \tilde{p}$$

as $\delta_n u \rightarrow 0$. Similar argument holds for $(f(2\delta_n u) - f(\delta_n u))/\delta_n u$ which is the directional derivative evaluated at $\theta_n^* + \delta_n u$.

Hence, the Hessian is constructed by the directional derivative. We need to obtain the numerical value of the directional derivative $\frac{\partial}{\partial \vec{u}} \log \tilde{p}$. Using the chain rule, a directional derivative $\frac{\partial}{\partial \theta} \log \tilde{p}_\theta$ can be expressed as $\partial_\theta(\lambda m) \times (\tilde{p}_\theta)^{-1}$ where $\partial_\theta(\lambda m) = \partial_\theta(\lambda_n(\theta)m(X, \theta))$

is a numerical derivative using the Romberg method¹⁵, see e.g. Korn et al. (2010).

The next task is to find a proper direction u . Because the direction u can be arbitrarily chosen¹⁶. We simply search the direction u using bisection method.

The bisection method concerns on $\tilde{\theta} = \theta_n^* + \delta_n u$ such that

$$\lambda(\theta) \sum_{i=1}^n m_i(\theta) = -\lambda_n(\theta) \sum_{i=1}^n m_i(\tilde{\theta}),$$

where $\sum_i m_i(\theta) = Z^T(Y - X\theta)$ in our experiment. So the simplified expression of $\tilde{\theta}$ is

$$\lambda(\theta)Z^T(Y - X\theta) = -\lambda(\theta)Z^T(Y - X\tilde{\theta})$$

or $X\tilde{\theta} = (2Z^TY - Z^TX\theta)$. Then the directional derivative $\frac{\partial}{\partial \tilde{u}} \log \tilde{p}$ can be set to $\partial_{\tilde{\theta}}(\lambda m)(\tilde{p}_{\tilde{\theta}})^{-1}$.

The Hessian used in the implementation is

$$\begin{aligned} \delta_n^{-2} K_{n,i,j} &\approx - \left[\frac{\partial}{\partial \tilde{u}} \left(\frac{\partial}{\partial \tilde{u}_1} \log \tilde{p} + \frac{\partial}{\partial \tilde{u}_2} \log \tilde{p} \right) \right] \\ &= \partial_{\tilde{\theta}}(\lambda m) \partial_{\theta}(\lambda m) \left[\frac{1}{(\tilde{p}_{\theta_n^*})^2} + \frac{1}{(\tilde{p}_{\tilde{\theta}})^2} \right] \partial_{\theta}(\lambda m) \partial_{\tilde{\theta}}(\lambda m). \end{aligned}$$

Step 3. After some rearrangement of $f(u)$, the linear term S_n can be expressed as:

$$\delta_n^{-1} S_n = \frac{3}{2} \frac{f(\delta_n u) - f(0)}{\delta_n u} - \frac{1}{2} \frac{f(2\delta_n u) - f(\delta_n u)}{\delta_n u},$$

which is a weighted average of numerical first derivative of $f(\tau)$ at $\tau = 0$ and $\tau = u$. We simply use $\partial_{\theta_n^*}(\lambda m)(\tilde{p}_{\theta_n^*})^{-1}$ to express $\delta_n^{-1} S_n$.

Step 4. Construct the adjusted estimator:

$$T_n = \theta_n^* + \delta_n K_n^{-1} S_n = \theta_n^* + (\delta_n^2 K_n^{-1}) \times (\delta_n^{-1} S_n)$$

¹⁵Because there is no closed form expression for λ , there is no way of obtaining analytical expression of $\partial_{\theta}(\lambda_n(\theta)m(X, \theta))$.

¹⁶The direction u_i and u_j are unknown. The directional derivative $\frac{\partial}{\partial \tilde{u}}(\cdot)$ depends on u_i and \underline{u}_j .

References

- Baggerly, K. A. (1998). Empirical likelihood as a goodness-of-fit measure. *Biometrika*, 85(3):535–547.
- Chan, K., Karolyi, A., Longstaff, F., and Sanders, A. (1992). An empirical comparison of alternative models of the short-term interest rate. *The Journal of Finance*, 47.
- Donald, S., Imbens, G. W., and Newey, W. (2003). Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Econometrica*, 117(1):55–93.
- Gnedenko, B. and Kolmogorov, A. (1968). *Limit Distributions for Sums of Independent Random Variables*. Addison-Wesley.
- Huber, P. (1981). *Robust Statistics*. Wiley, New York.
- Kallenberg, O. (2002). *Foundations of Modern Probability*. Springer Press.
- Kitamura, Y., Otsu, T., and Evdokimov, K. (2009). Robustness, infinitesimal, neighborhoods, and moment restrictions. *forthcoming in Econometrica*.
- Kitamura, Y. and Stutzer, M. (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, 65(4):861–874.
- Kitamura, Y., Tripathi, G., and Ahn, H. (2004). Empirical likelihood-based inference in conditional moment restriction models. *Econometrica*, 72(6):1667–1714.
- Korn, R., Korn, E., and Kroisandt, G. (2010). *Monte Carlo Methods and Models in Finance and Insurance*. CRC Press.
- Le Cam, L. (1974). *Notes on Asymptotic Methods in Statistical Decision Theory*. Centre de recherches mathématiques, Université de Montréal.
- Le Cam, L. and Yang, G. (1990). *Asymptotics in Statistics: Some Basic Concepts (Springer Series in Statistics)*. Springer-Verlag, New York.
- Le Cam, L. and Yang, G. (2000). *Asymptotics in Statistics: Some Basic Concepts Second Edition (Springer Series in Statistics)*. Springer-Verlag, New York.

- Newey, W. and Smith, R. J. (2004). Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255.
- Owen, A. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1):90–120.
- Owen, A. (2001). *Empirical Likelihood*. Chapman & Hall/CRC, Florida.
- Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22(1):300–325.
- Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *The Annals of Statistics*, 35(2):634–672.
- Smith, R. (2005). Local gel methods for conditional moment restrictions. Technical Report CWP15/05.
- Smith, R. J. (1997). Alternative semi-parametric likelihood approaches to generalised method of moments estimation. *The Economic Journal*, 107(441):503–519.
- van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.
- Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, 20(4):595–601.